

## Free and public expression

**After a slow start, progress towards developing public repositories for gene expression data is poised to accelerate. For the many biologists working with DNA microarrays, that should be welcome news.**

In the era of functional genomics, DNA microarrays have emerged as uniquely powerful tools for studying gene expression. They allow biologists to take samples of tissue and study the activity of thousands of genes at a time. Such experiments churn out reams of data, of which only a concise summary appears in published papers.

Given some of the potential pitfalls of microarray research (see page 860), it is important for other researchers to have access to the underlying data. But even for the most carefully conducted studies, anyone wanting to build from previous work would be foolish not to scrutinize their peers' data before designing their own experiments.

It is for this reason that *Nature* requires microarray data relating to papers we publish to be made freely available. But at present, the system for doing so is far from perfect: data are generally posted on authors' own websites; there is little standardization, and no firm guarantee that data will remain available in perpetuity.

It would be much better for the data to be submitted to public repositories along the lines of the GenBank sequence database. But developing a robust and useful database for gene expression data has proved a more formidable challenge.

DNA sequences are linear strings of data, and require relatively little ancillary information — such as the species involved, and whether the sequence comes from nuclear DNA, or was derived from messenger RNA. Gene expression data, on the other hand, are richer and much more varied. Readouts from DNA microarrays also make sense only in the light of a suite of supporting information defining the type of chip used, the nature of the tissue samples involved, the precise procedures used to hybridize the RNA in the samples to the DNA spots on the chip, and so on.

These challenges may explain why public repositories for gene expression data have been slow to get off the ground. Now, however, several embryonic databases exist, notably the Gene Expression

Omnibus (GEO) at the National Center for Biotechnology Information in Bethesda, Maryland (<http://www.ncbi.nlm.nih.gov/geo/>), and ArrayExpress, run by the European Bioinformatics Institute at Hinxton near Cambridge (<http://www.ebi.ac.uk/arrayexpress>). A new consortium is also proposing to construct a database for gene expression in cancer and other diseases (see page 855).

So far, existing databases have attracted few submissions. But that may soon change, as bioinformaticians are close to agreeing a standard mark-up language for formatting gene expression data. There have been two competing standards, MAML and GEML. Proponents of the two are now working to adopt a unified standard, incorporating aspects of both. By mid-June, this is expected to be adopted by an expert panel known as the Object Management Group.

With a single format for gene expression data, databases should be able to 'talk' to one another and exchange data. The existence of a standard language should also spur development of software tools to query the databases, and to manage and display gene expression data.

Issues remain to be resolved, however. Perhaps the most important is agreement on the suite of supporting information to be included in the databases to allow entries to be readily interpreted. The international Microarray Gene Expression Database Group (<http://www.mged.org>) has applied considerable thought to this question, developing guidelines called MIAME (for Minimal Information About a Microarray Experiment). But they are not universally accepted, and at present GEO seems to be less concerned than ArrayExpress about data being MIAME-compliant.

*Nature* and its sister journals will be watching developments closely, to determine the point at which it may become appropriate to require authors to submit microarray data to a public database or databases. In the meantime, we welcome the views of the diverse community of biologists interested in gene expression. ■

## Science without a tinge of green

**With no explanation, the first budget proposed by George W. Bush proposes drastic reductions in environmental research.**

The Bush administration has tried to portray its 2002 budget proposal as a case study in moderation. But when it gets down to the small print, the proposal is anything but moderate, slashing programmes that are not to the administration's liking.

Environmental science is one of the victims. Research that deals with air and water pollution, or with biodiversity, would be cut to the bone under the proposal. Practically no agency that supports work in toxicology, hydrology, oceanography, atmospheric science or whole-organism biology is left untouched.

The numbers are, by the usual standards of such things, quite remarkable. The research programmes of the Environmental Protection Agency, for example, are reduced by more than 20%, from \$919 million to \$707 million; biological and environmental science at the Department of Energy falls by 8%; and the United States Geological Survey's water-resources division is cut by almost a quarter.

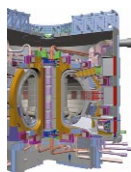
The National Science Foundation may have a level budget for

research, but this means that the agency's ambitious plans to expand its environmental-science programmes will be stillborn. The science office at the National Oceanic and Atmospheric Administration, meanwhile, loses 10% of its funding. Even the Smithsonian Institution, sensing that the new administration is not going to be a generous benefactor for environmental research, has seen fit to close an important conservation centre (see *Nature* **410**, 727; 2001).

There has been no strategic explanation for any of this. One is left to conclude that the Bush administration — which has not yet named appointees to its most important scientific positions — simply does not buy the case for the scientific study of environmental questions.

Yet an unusually wide consensus in the United States holds that good science is the only reasonable basis for environmental stewardship. A majority of members of both houses of Congress claim to believe that they should act now to reverse Bush's rash attack on environmental science. ■





**Slow progress**  
Japan edges towards  
hosting experimental  
fusion reactor  
p856



**Testing times**  
German animal  
experiments targeted  
by campaigners  
p857



**Plain sailing?**  
Potential new mode  
of space transport  
suffers setback  
p858



**News feature:**  
DNA microarrays  
Hot property or lab  
liability?  
p860

# Astronomers concerned by plan to give NASA ground control

**Tony Reichhardt, Washington**

A White House proposal to consider merging space-based astronomy funded by NASA with ground-based astronomy funded by the National Space Foundation (NSF) is alarming US astronomers. But they concede that without some such radical change their funding may be at risk.

As part of its 2002 budget request, the Bush administration called for a panel of experts to investigate "the pros and cons of transferring NSF's astronomy responsibilities to NASA". The panel — whose members will be named this week — is free to suggest other options for managing the government's astronomy portfolio. Its investigation will be overseen by the National Academy of Sciences and should be completed by September.

NASA and NSF provide more than 90% of all funding for US astronomy. But NASA's share is much larger because of the high cost of putting instruments such as the Hubble Space Telescope and the Chandra X-Ray Observatory into orbit. Indeed, NSF's budget request for astronomy next year is \$156 million, compared with the space agency's \$1 billion. NASA has also recently eclipsed NSF in becoming the primary sponsor of individ-



Pair of eyes: the Hubble Space Telescope and (inset) the Gemini South telescope, Cerro Pachon, Chile.

ual investigator grants in astronomy — twenty years ago, most came from NSF, but now three-quarters are from NASA.

NSF has taken on an increasing number of large commitments, such as the twin Gemini telescopes in Hawaii and Chile. And new projects recommended by the latest 'decadal review' for astronomy (see *Nature* 405, 381–382; 2000), whose ground-based component alone is expected to cost \$1 bil-

lion over the next ten years, could strain NSF's astronomy budget to breaking point.

Giving NASA, with its greater budget and experience with big-ticket projects, responsibility for all US astronomy might appear to make sense. But a discussion at a meeting of the academy's Committee on Astronomy and Astrophysics last week suggested that astronomers will resist the idea.

Astronomers at the meeting questioned whether the two agencies' philosophies and management styles could be matched. NASA is a mission-orientated agency focused primarily on launching hardware, whereas NSF takes its cues from the research community. "Science does not always come in as the number one priority" at NASA, said University of Wisconsin astronomer Blair Savage.

One option would allow NASA to handle the construction of all large facilities and leave investigator grants to NSF. But this might marginalize astronomy at NSF, warned Ellen Zweibel of the University of Colorado.

Committee co-chair Richard McCray of the Joint Institute for Laboratory Astrophysics in Colorado also saw value in having two competing agencies. But NASA's big budget may tempt ground-based astronomers used to making do with less. A project such as the Atacama Large Millimeter Array (ALMA) "is going to stretch the capacity of NSF in a big way", said McCray, but "would probably get done a lot faster" at NASA.

## EU plans global positioning system

**Jim Giles, London**

The European Union (EU) is planning a satellite-based global navigation system that will rival the US-operated Global Positioning System (GPS).

Transport ministers from EU states agreed last week to spend 100 million euros (\$88 million) on developing the new system, called Galileo. The EU and European Space Agency (ESA) are expected to release a further 1 billion euros later this year. Another 2.1 billion euros will be needed for the system, which will involve launching at least 20 satellites between 2005 and 2008.

Europe aims to depend less on GPS for transportation, science and other purposes, partly because the US military restricts

access to its highest-precision capabilities.

Jim Davis, a geologist at the Harvard-Smithsonian Center for Astrophysics in Cambridge, Massachusetts, who uses GPS, says the European system will be welcomed if it delivers more resolution than GPS.

"Scientists need more accurate knowledge about the position of the satellites than the military does," he says.

Galileo could provide quick, precise measurements. "Its real difference is speed," says Hans Fromm, head of navigation at ESTEC, an ESA laboratory in Noordwijk, the Netherlands. "It takes hours to get an accurate fix from GPS. Galileo should be able to do it in ten minutes or so." Researchers are expected to have free access to the system. ■



## Earth sciences and biodiversity fall victim to budget plan

Rex Dalton, San Diego

Two major collaborative projects costing some \$175 million are among the casualties of the budget proposal for 2002 put forward this month by the National Science Foundation (NSF).

EarthScope, a comprehensive study of North America's lithosphere and crust, and the National Ecological Observatory Network (NEON), a proposal for long-term biodiversity monitoring, are among major facilities which the NSF left out of its request.

Agency officials say the lack of any extra research funding in the \$4.5 billion NSF budget left no room to make a start on either project next year. Researchers involved in years of planning for the projects, which had been approved by the National Science Board, the NSF's governing body, expressed bitter disappointment at the news.

"The community can get exhausted if it has to keep up this level of activity without any scientific return," says Mark Simons, a geophysicist at the California Institute of Technology who was involved in planning a component of EarthScope.

"It probably should come as no surprise," says James Beach, an informaticist at the University of Kansas involved in the planning of NEON. "Why build a national infrastructure to monitor long-term environmental trends, when the Bush administration [wants] to accelerate the private plunder of the country's natural capital?"

Some US researchers are now looking to an international effort to advance a component of EarthScope that would involve drilling 4 kilometres into the San Andreas fault near San Francisco to conduct seismology tests. Mark Zoback, a seismologist at Stanford University who helped plan this project, says that the International Continental Scientific Drilling Program at GeoForschungs-Zentrum in Potsdam, Germany, is considering support for a pilot drill hole 2 km deep. A decision is expected soon.

But such activity is a far cry from that envisaged under the original \$75 million Earthscope proposal (see *Nature* 405, 390–392; 2000), which also included a satellite-based radar system and an array of 2,400 mobile seismometers deployed across the United States.

For NEON (see *Nature* 404, 216; 2000), scientists had sought \$100 million to build and equip 10 environmental monitoring sites. ■

## Britain looks to wave-power site after researchers flee

David Adam, London

Britain is considering allocating a stretch of coast to research into wave and tidal energy, after planning regulations forced an emerging tidal-power company to flee to Iceland.

The new site would allow tidal and wave power to be assessed side by side and feed electricity into the national supply. The Scottish parliament has already suggested locating the site off the west coast of Scotland.

Britain's location and rugged coastline make it ideal for such research, but researchers say they have to obtain written planning consent from up to 14 different organizations.

"The only problem we've got [in Britain] is the lack of long-term planning," says John Hassard, co-founder of tidal-energy company RV Power. "We could have got permission for our installation in the Solent [off the south coast of England] but it would have taken between 18 months and two years."

The paperwork included forms from the coastguard, port authority, countryside groups, two local councils and three government departments.

The company will now install prototype equipment off the coast of Iceland. "We made one phone call," Hassard says. "A lady came down in a Land Rover, met us by the water and said 'Go ahead and build it. How



Iceland's coast proves hospitable to researchers.

can we help?'" Iceland is enthusiastic about renewable energy — geothermal sources satisfy much of its electricity demand.

Only a few such projects are scattered around Britain's coastline. "The problem we had was knowing where to start and who to talk to," says Charles Taylor, senior engineer with wave-power company Ocean Power Delivery in Edinburgh. He says Britain needs to improve its planning procedures.

UK energy minister Peter Hain told a recent parliamentary committee hearing: "There is a serious problem here, both in the sheer bureaucracy and in the associated costs." The committee's report on wave and tidal power is expected to appear in a few weeks, and to recommend the creation of a dedicated research site. ■

## Search is on for underwater volcano

Rex Dalton, San Diego

A team of oceanographers is on a mission to track down and monitor an underwater volcano that has erupted about 300 kilometres off the coast of northern California.

The research vessel *New Horizon* sailed this month from its base at the Scripps Institution of Oceanography in La Jolla, California, to arrive at the suspected site of the volcano last week. So far, two days of attempts to find the Gorda Ridge's linear volcano — the eruptions of which were first detected by military sonar — have proved unsuccessful.

"It's like fishing, but a lot more expensive," says Ed Baker of the Pacific Marine Environmental Laboratory in Seattle, Washington, who is jointly leading the expedition.

The team plans to spend a week casting instruments 3,200 metres beneath the ocean surface to measure temperature and chemical composition, with the aim of locating the volcano's plume. The faster the vessel reaches the area, the better chance it has of finding the plume before it dissipates.

Oceanographers are seeking better equipment for examining volcanic events, as part of an international effort known as Dynamics of Earth and Ocean Systems (DEOS). This would include seismometers on the sea floor off the US Pacific coast.

DEOS has attracted some UK funding. But hopes of support from the US National Science Foundation have been hit by research funding restraints (see story, left). ■



New lava on old: one sign of volcanic activity.



# Cancer comes under scrutiny in fresh genomics initiative

Jonathan Knight, San Francisco

Genome researchers are planning an ambitious project to catalogue all the changes in gene activity associated with any type of cancer.

Organizers of the project, dubbed the International Genomics Consortium (IGC), say that ultimately it will collect gene expression information on a wide variety of diseases.

But some researchers fear that the project will duplicate existing efforts. And they worry that it might impose a single standard at too early a stage in the development of the DNA chip technology used to gather information on gene activity.

Biomedical researchers are increasingly dependent on DNA chips, also known as microarrays, because they can monitor changes in the expression of thousands of genes at once (see News Feature, page 860). Cancer researchers, for example, use chips to spot abnormal gene activity in tumours and so identify possible drug targets.

The project's \$45 million cost would be paid by the manufacturers of microarrays and associated technologies, but all data would be publicly available without restriction, says organizer Andy Baxevanis, director of computational genomics at the National Human Genome Research Institute (NHGRI) in Bethesda, Maryland.

Although both Baxevanis and co-organizer Jeff Trent are NHGRI researchers, they say that the project can proceed without federal support. They are currently negotiating with potential academic partners and industrial sponsors. Although no institution has officially signed up yet, dozens have expressed interest, Baxevanis says, including several in Europe.

In its first phase, the IGC would seek to assemble microarray data on 10,000 tumour samples from cancer centres worldwide by the middle of next year.

The IGC database would link the tumour gene profiles with anonymous clinical data on the patient. This would give scientists a quick way to see if a gene activated in breast cancer, for example, is also activated in other types of cancer. And oncologists would have immediate access to clinical records of patients with tumours that genetically match those of their own patients.

Rather than compiling existing gene expression data, the IGC would create its own, so that it could all be properly integrated into one database. Currently, different laboratories use different microarrays with different sets of genes, making database integration problematic (see page 851).

The organizers say they intend to work



Growing up: the chips are down for tumours, with plans for a database of the genes behind cancer.

with DNA chip companies such as Affymetrix and Agilent Technologies, both of northern California, to develop one or two new standardized microarrays. Each participating institution would use the same standard chips under the supervision of IGC representatives who would ensure that all samples are treated the same way.

But some cancer researchers fear that the proposed project would merely duplicate their existing efforts. Samir Hanash, whose lab at the University of Michigan is one of about 20 funded by the National Cancer Institute to do expression profiling of cancer, says it would be better to combine all microarray data in a single location, rather than creating a separate database.

The more experiments are compared, the easier it will be to draw meaningful conclusions about the genes involved in cancer, he says. Hanash fears that the new collaboration will divide the available information between existing databases and the new one to be created. "There is a danger the effort will be diluted," he says.

Additionally, Hanash suggests, it may be too early to decide on the right gene-profiling standards. DNA chips measure the levels of messenger RNA, the direct product of genes, but proteins may be a better measure of gene activity, and could potentially be studied directly by new technologies. "Maybe microarrays are in fashion today but they still represent just one technology platform," he says.

But Baxevanis argues that establishing a standard for DNA chips now is the best way forward. "One reason expression databases have difficulty getting off the ground is they are comparing apples and oranges," he says. ■

## Blood product from cattle wins approval for use in humans

Corie Lok

An animal-derived blood substitute has been approved for use in humans in South Africa.

Hemopure, an oxygen-carrying compound derived from bovine haemoglobin, has been given the go-ahead for treating acute anaemia and for use during surgery.

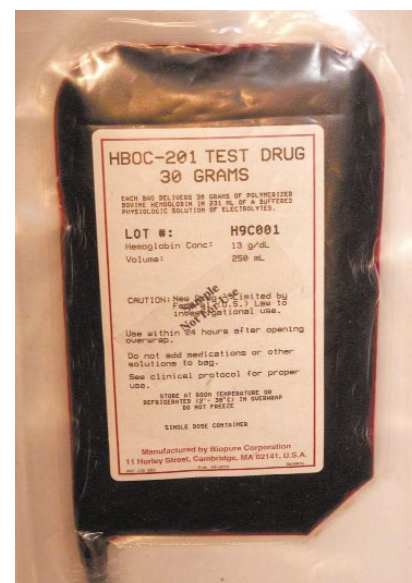
Luc Noel, head of the World Health Organization's blood safety unit, says that treatment could be useful in South Africa's rural areas where safe blood is in short supply.

The World Health Organization estimates that one in five adult South Africans are infected with HIV. Noel says that he welcomes the new product if it reduces reliance on blood transfusions.

Hemopure was developed by Biopure of Cambridge, Massachusetts. The company says that the product is compatible with any blood type.

The raw haemoglobin used in Hemopure is harvested from US beef cattle bound for slaughter. Biopure requires the farmers to keep records of the cows' origin, feed, medical history and condition. The company says that its purification process removes infectious agents such as HIV, hepatitis C and transmissible spongiform encephalopathy agents.

Biopure says it will seek approval for Hemopure in the United States later this year, and then in Europe. ■



In the red: a blood product from animals could reduce the need for blood transfusions.

# Japan prepares to pitch for international fusion reactor

David Cyranoski, Tokyo

Japan has taken another tentative step towards hosting the International Thermonuclear Experimental Reactor (ITER), a collaborative project to develop nuclear fusion as an energy source.

Earlier this month, the country's Atomic Energy Commission approved a report supporting construction of the multibillion dollar project in Japan. But the commission also emphasized that hosting the facility must not harm existing Japanese research programmes.

The commission signalled its intention to bid for the ITER late last year, and its report endorses the project's technological feasibility, safety and scientific importance. The plans will probably be submitted early next month to Japan's top scientific panel, the Council for Science and Technology Policy (CSTP), which is chaired by the prime minister.

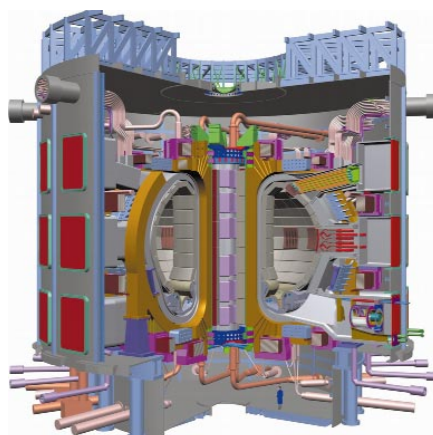
If the council approves the plan, Japan will select a suitable site and put it forward in time for the mid-2001 deadline set by the ITER consortium. Other possible candidate sites for the experiment are in Canada and France, although neither country is as well placed as Japan to shoulder the two-thirds of construction costs likely to fall on the host.

A lack of natural energy resources makes Japan the obvious place for the project, says Hiroyuki Yoshikawa, president of the National Institute of Advanced Industrial Science and Technology and chair of the panel that drew up the commission's report. "Japan's energy future is with fusion," he says.

But there are already several fusion projects under way in Japan, which researchers are loath to see cut back. The Japan Atomic Energy Research Institute (JAERI) and the National Institute for Fusion Science (NIFS) are both investigating doughnut-shaped plasma devices that would magnetically confine plasma and heat it to the temperatures at which fusion would occur.

The JAERI is working with 'tokamak' reactors, of which the ITER would be the largest ever built, whereas NIFS is concentrating on helical devices, which use a different technique to create and sustain their magnetic fields. Another approach under investigation, called inertial confinement fusion, relies on a powerful laser to create fusion in a tiny pellet of fuel.

Competition between the approaches has intensified since they all fell under control of the education ministry, as a result of January's government reorganization (see *Nature* 408, 757; 2000).



Getting there: Japan's plan to host ITER may be endorsed by its top science panel next month.

"Scientifically speaking, [the ITER] is clearly something we should do, but splitting up the budget has been a problem," says Tatsuhiro Yamanaka, director of Osaka University's laser fusion research centre, home of Japan's main inertial fusion programme.

As a result, Yoshikawa's report backs ITER construction only if funding for all of the other fusion projects is maintained.

Some believe that Japan will be able to do that. "Japan will put money where it needs to," says Masato Nakamura, director of the government's fusion energy office. The estimated ITER budget of ¥400 billion (US\$3.3 billion) would fit easily in Japan's plan to spend an average of ¥4.8 trillion a year on research and development, says Yoshikawa.

But others are less certain. "There is still no consensus within Japan about how to run such a project," says Atsuo Iiyoshi, president of Chubu University in Nagoya, a former director of NIFS and a member of the panel. He says that extra materials research needed to make the ITER work could cost ¥100 billion, and he questions Japan's ability to coordinate the project effectively.

Iiyoshi says the commission's report reflects some of this caution, and stops short of a full-scale endorsement of bringing the project to Japan. "It's basically just passed the buck to the CSTP," says Iiyoshi.

But there is plenty of enthusiasm for fusion power at higher levels of government in Japan. The country will pursue it as the "ultimate" energy source, says Hiroo Imura, a former president of Kyoto University who sits on the CSTP. He says: "The council will have to consider carefully the budget allocation to various types of nuclear fusion research when deciding whether to support the ITER."

## Swiss proteomics company aims to make big impact

Alison Abbott, Munich

Swiss proteomics pioneers hope to make a big splash in the commercial world when their new company, GeneProt, opens its first facility in Geneva on 26 April.

The company hopes to become one of the main players in proteomics, having raised US\$122 million in capital since it was founded a year ago. It is now building a second facility at Princeton, New Jersey.

GeneProt will employ 100 people, and aims to discover proteins that could provide targets for drug development or diagnostics, as well as ones that could be used directly as therapeutics. Its researchers will separate tens of thousands of proteins from diseased and healthy tissues, using gel filtration and chromatography, and identify them using a bank of 51 mass spectrometers.

Analysis will be supported by a Compaq computer, which Denis Hochstrasser, one of GeneProt's founders, says will deliver more power than is available to any other proteomics company.

Hochstrasser and the other founders of GeneProt have been making Geneva a centre of proteomics research since the 1980s. He is a vocal proponent of free public access to genomic and protein information, and says that GeneProt will make information publicly available about the "many proteins, such as housekeeping proteins, which have no commercial value".



Free thinker: Hochstrasser says he will make much of GeneProt's data freely available.

ITER

GENEPROT



# Satellite will probe mutating seeds in space

David Cyranoski, Tokyo

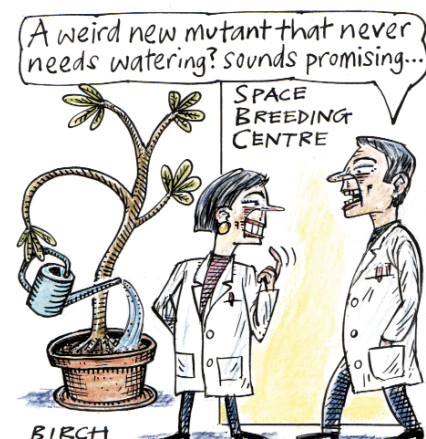
Chinese plant scientists are hoping that a proposed new satellite will deliver not only agricultural breakthroughs but also a greater understanding of how to make plants mutate.

The satellite project, which was announced late last month, is designed to expose plant seeds to cosmic radiation, zero gravity and other effects that are encountered only in space.

It aims to "make breakthrough mutations with important economic significance, such as crop yield and quality", according to Liu Luxiang, director of the Space Breeding Centre at the Chinese Academy of Agricultural Sciences, which is promoting the project.

Chinese scientists say they have already discovered several rare, inheritable genetic mutations — capable, they say, of producing giant sweet peppers as well as improving the quality of rice and wheat — using recoverable satellites and high-altitude balloons.

The new project aims not only to identify more mutants, but also to establish their causes. "It could take advantage of the many kinds of radiation, such as gamma rays, X-rays and secondary emissions, at many ener-



gy levels, that can affect plants and seeds," says Didier Schmitt of the European Space Research and Technology Centre (ESTEC), a European Space Agency laboratory in Noordwijk, the Netherlands, which is not involved in the project.

The satellite will be designed with minimum protection against radiation — in contrast to the heavily protected research facilities sent into space by Russia and the United States. Radiation detectors and special com-

partments — such as one protected against radiation and another containing a gravity simulator — will help researchers to identify the role of these factors in causing mutations.

The fact that cosmic radiation causes mutations is well known, but the effect of microgravity remains disputed. Some researchers think that the combination of the two could produce some interesting results. "There might very well be a synergy effect of radiation and lack of gravity," says Enno Brinckmann, ESTEC's senior biologist.

Seeds will be loaded onto the satellite, which will then be launched into orbit. When the satellite is recovered, the seeds will be grown and screened for mutants in a process that will take three to five years.

The plan, which could cost as much as RMB200 million (US\$25 million) for the satellite alone, has yet to be approved by the Chinese government, but Liu says he is optimistic that it will go ahead.

The project "would be unique as an unattended facility specifically for plant mutation studies," says Brinckmann. Schmitt agrees that it would complement other satellite-based biological research as well as experiments planned for the International Space Station. ■

# Animal labs fearful over activists' plan to name names

Quirin Schiermeier and Alison Abbott, Munich

Biomedical research institutes in Germany are on the alert this week over plans by animal rights advocates to publish the names and work addresses of scientists who use laboratory animals.

The 400-page list will be published on 24 April as part of a campaign by the Bundesverband der Tierversuchsgegner (BVTVG), an organization that opposes any form of animal experimentation.

BVTVG produced its list by systematically analysing 1,800 scientific papers published in 1998 and 1999. "We will point the finger at this hidden side of research," says Corina Gericke, a spokeswoman for the group.

Today's animal rights movement in Germany is professionally organized, and financially strong. But there have been few violent incidents, and only occasional cases of harassment. Most protests have concerned animal transportation and the fur trade, rather than research.

Research groups fear this may be about to change. "There are clear signs that protests are switching to animal experimentation," says Ivar Aune, spokesman for the research lobby group Society for Health and Research.

Last month, for example, more than 1,000

participants at a physiologists' meeting in Berlin were evacuated after a bomb threat by an anonymous animal rights group.

Aune fears that militant activism could increase in the wake of this latest personalized campaign. He says research institutes should increase safety measures.

The University of Düsseldorf, which three years ago was stung by a disinformation

campaign by animal activists, is planning its own counter-attack. "It is important for the public to understand that all animal experiments are carried out in line with strict German animal protection laws, and only after all alternative methods have been explored and excluded," says Annemarie Treiber, president of Germany's Society for Laboratory Animal Science, and the official responsible for animal welfare at the university.

Three years ago, neuroscientists complained that large German research organizations were too passive in their response to animal rights activists (see *Nature* 396, 505; 1998).

But Kuno Kirschfeld, who deals with animal welfare at the Max Planck Society, one of Germany's largest research organizations, points out that the society cannot intervene unless there is illegal action or incitement to use violence.

The society has set up an office to deal with questions from the public about biological medicine, including animal experimentation. Its head is Jan Erik Bohling, a former spokesman for the Society for Health and Research. Bohling says he favours openness with the public and dialogue with groups such as BVTVG. ■



'Pointing the finger': researchers will be named.



### Pharmaceutical firm wins access to vast Chinese DNA library

**London** Pharmaceutical company AstraZeneca has signed a deal with Shanghai's Jiao Tong University giving it access to the university's library of DNA samples collected from thousands of schizophrenics.

China is thought to offer a particularly rich seam for genetics research because the population is relatively homogenous, which makes it easier to spot small genetic mutations associated with different diseases (see *Nature* **410**, 10; 2001). AstraZeneca claims that its new project is the largest research programme yet to make use of Chinese DNA.

The university's collection of blood and tissue samples includes some from extended and nuclear families where schizophrenia was prevalent, as well as control samples from unrelated patients.

### UK sets up gene bank to protect rare sheep breeds

**London** A gene bank to protect rare sheep breeds facing extinction because of Britain's foot-and-mouth epidemic has been launched by scientists at York and Leeds universities.

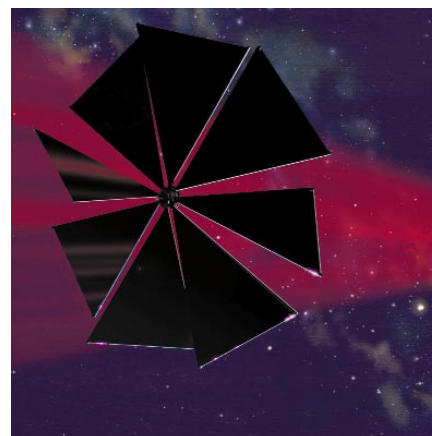
The Heritage GeneBank will collect and freeze semen, eggs and embryos to protect the identities and diversity of British sheep breeds. The scheme will initially concentrate on the Herdwick breed, which is unique to Britain and found mainly in the Lake District in the north of England. About a quarter of the existing 100,000 Herdwick sheep have already been slaughtered. The gene bank also plans to cover other rare and speciality breeds that are under threat.

### Solar sail launch put back as rocket fails

**Washington** Plans for the first ever deployment of a solar sail in space suffered a setback last week when the rocket intended to launch it was damaged during ground testing.

Solar sails are large sheets of material that can, in theory, use radiation from the Sun to propel spacecraft in much the same way as conventional sails use wind to drive ships. Known as Cosmos 1, the 600-metre-square sail was developed by the US-based Planetary Society, a non-profit-making organization dedicated to exploring the Solar system. It was due to be launched on 26 April from a Russian nuclear submarine in the Barents Sea. The society still hopes to launch the sail later this year.

The solar-sail programme has had many



Setting sail: Cosmos 1 was designed to propel spacecraft by harnessing solar power.

critics in the scientific community. Detractors say that the acceleration of the sails would be too slow for them ever to be used for manned space flight.

### Malaria-parasite genome database goes up on web

**Paris** Malaria researchers will be able to access the genome database for the malaria parasite *Plasmodium falciparum* free of charge on the web for the first time, as the result of work done by an international

collaboration led by the University of Pennsylvania.

Individual partners were already running their own databases, but their information has now been compiled in a form that will be easier to use. Specially developed data-mining tools will allow easy searching of the raw data on the parasite's 14 chromosomes, the collaboration says.

Sequencing material for the database was provided by the Institute for Genomic Research and the Naval Medical Research Centre in Maryland, Stanford University in California, and Britain's Sanger Centre.

♦ <http://plasmodb.org>

## Alternative clinical centre names head

**Washington** Marc Blackman, professor of medicine at the Johns Hopkins University School of Medicine, has been appointed head of the new clinical branch of the US National Center for Complementary and Alternative Medicine (NCCAM).

The NCCAM was set up by the National Institutes of Health (NIH) in 1998 to assess alternative and complementary medical therapies. The new branch, to be based on the NIH campus in Bethesda, Maryland, will extend the centre's work by conducting clinical trials of potential treatments.

## Gene-therapy trial for Alzheimer's underway

**San Diego** The first clinical trial to use gene therapy to treat a neurodegenerative disease is being carried out at the University of California, San Diego (UCSD).

Earlier this month, skin cells that had been genetically modified to produce nerve-growth factor were implanted into the brain of a 60-year-old woman with Alzheimer's disease. Researchers Mark Tuszynski of UCSD and Fred Gage of the Salk Institute in La Jolla, California, placed the cells in an area of the brain called the nucleus basalis. Patients with Alzheimer's are known to lose cells from this area.

Some experts are concerned that the implanted cells have not been engineered to contain 'suicide' genes, which would allow them to be destroyed if they cause unwanted side effects. A recent trial involving the use of fetal brain cells to treat Parkinson's disease has raised similar concerns (see *Nature* 410, 401; 2001).

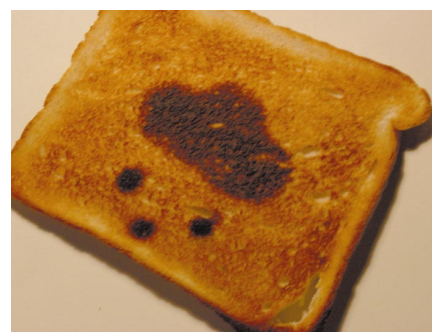
## Toast the Internet over breakfast

**London** Decisive proof that the Internet really is changing our lives has arrived, with the development of a toaster that can link to the

World-Wide Web and imprint the weather forecast onto your breakfast.

The toaster plugs into the Internet using a normal telephone connection and downloads the day's forecast from a specially designed website. The bread is then toasted normally until the final 20–30 seconds, when a stencil in the shape of sun, rain or cloud rolls out over the heating element. The forecast is visible as a dark pattern on the toast.

Robin Southgate, a design student at Brunel University in west London, developed the toaster in conjunction with Stan Swallow of the university's Design for Life centre. He hopes to use more flexible stencils to display text messages or adverts on toast.



Your toast: a stencil allows images from the web to be imprinted onto your daily bread.

# When the chips are down

DNA microarrays are transforming studies of gene expression. But some of the biologists flocking to exploit this powerful technology are not aware of its potential pitfalls. Jonathan Knight relates a cautionary tale.

Some call them DNA chips, others microarrays, but whatever name you prefer, they are one of the hottest tools in biology. A search of the Medline database for papers published in 1999 with 'microarray' in their title yields just 27 results. Try the same search for 2000 and the number jumps to 97 — a crude measure, perhaps, but it is a testament to a revolution that is transforming studies of gene expression. As the genomics revolution begins to make its mark, biologists are turning in growing numbers to a technology that lets them analyse cells or tissues and determine, at a stroke, which genes are active.

DNA microarrays consist of a library of genes immobilized in a grid, usually on a glass slide. Each individual 'spot' in the grid contains DNA from a single gene that will bind to the messenger RNA (mRNA) produced by the gene concerned. So by liquidizing a sample from a given tissue type, tagging its mRNAs with fluorescent dyes and then exposing the sample to the slide, it is possible to obtain an instant visual read-out revealing which genes were active.

Researchers who previously studied the activity of one gene at a time can now analyse the expression of thousands of genes simultaneously. But as aficionados explore the technology's limits, they are turning up errors in DNA chips that could lead unwary biologists towards erroneous conclusions. And experts worry that too few of the

researchers rushing to embrace DNA microarrays are aware of the potential pitfalls. "It's going to revolutionize science. But the technology is in its infancy, so there are going to be some growing pains," says Timothy Zacharewski, a toxicologist at Michigan State University in East Lansing, who makes and uses microarrays. "It's amazing how many people are going forward without a full appreciation of what they are getting into."

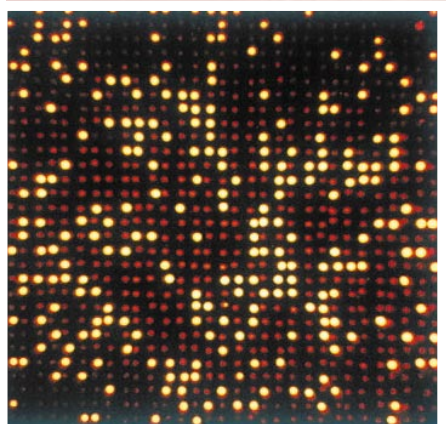
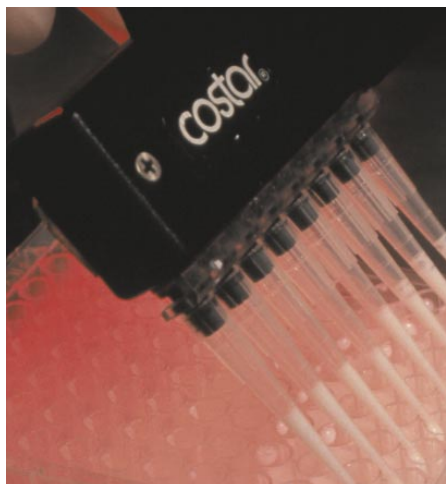
The enormous number of genes that can be studied at one go is the technology's curse, as well as the source of its power. Although microarray production is heavily automated, there are many opportunities for human error. "For any experiment you can mislabel a tube and mess yourself up," says Joseph DeRisi, a microarray pioneer at the University of California, San Francisco. "But here, the potential for the error to magnify itself is much more drastic. Instead of one tube at a time, you are doing 6,000."

## Send in the clones

One popular type of array was devised by a team led by Patrick Brown at Stanford University in California<sup>1</sup>, and is based on libraries of gene sequences made using mRNA. To store and reproduce these sequences, researchers make 'complementary' DNA (cDNA) copies of the RNA messages and splice them into loops of DNA called plasmids. The plasmids are then inserted into bacteria, which grow in cul-



ANTHONY PIDGEON



Power tools: microarrays (left) show quickly and easily which genes in a sample are active, but despite automated steps in their manufacture, the chips are still open to significant errors.

ture and churn out more plasmids from which the cDNAs can be derived for spotting onto microarray slides.

Errors creep in as these bacterial cultures, or the cDNA clones extracted from them, are manipulated. The cultures are often stored in small plastic plates, each typically containing 96 wells, and they are transferred from plate to plate using pipetting robots. But bacteria can easily contaminate other wells, and technicians can make errors such as loading plates into the robots the wrong way round or taking samples from the wrong well for sequencing. As a result, between 1% and 5% of the clones in even the best-maintained sets do not contain the sequence that they are supposed to.

Until recently, few researchers were aware of the extent to which the errors can multiply as clone sets are copied and transferred from lab to lab. But last year, after hearing anecdotal reports of high error rates in a set of mouse cDNA clones assembled by a group of labs called the IMAGE (Integrated Molecular Analysis of Genomes and their Expression) consortium, Zacharewski decided to investigate further.

The IMAGE consortium has compiled a variety of cDNA clone sets, which are now produced by commercial suppliers. Scientists wanting to use IMAGE clone sets for microarray studies can either buy bacterial cultures or purified cDNAs and make up their own slides, or order pre-manufactured chips.

To check the accuracy of commercially available IMAGE mouse cDNA clone sets, Zacharewski and his colleagues purchased a



**I**t cannot be assumed that microarrays based on cDNA clones are reliable.

set from one supplier, Research Genetics of Huntsville, Alabama, and sequenced 1,189 cDNAs. Only 62% of the stocks definitely represented a pure sample of the correct clone<sup>2</sup>. Of the remainder, more than half seemed to contain the wrong cDNA, and the rest contained either a mix of different cDNAs or did not yield a readable sequence.

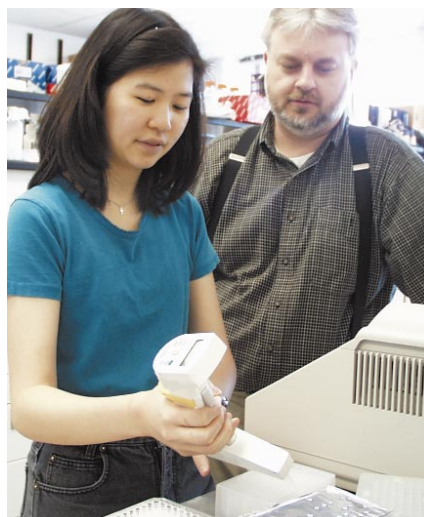
In some cases, the apparent errors may mean that the sequence for the clone deposited in the public databases is wrong, rather than there being a problem with the clone. But stocks containing more than one cDNA were probably the result of cross-contamination, Zacharewski says. Other problems may reflect handling errors accumulated as different labs managed and distributed the stocks over the years.

Before Zacharewski's study, reagent suppliers had acknowledged the potential for errors and started producing cleaned-up, 'sequence-verified' cDNA clone sets. But even these can be problematical. Indeed, researchers at three major microarray centres told *Nature* that they have found disturbingly high error rates — up to 30% — in copies of the sequence-verified version of the Research Genetics mouse cDNA clone set studied by Zacharewski's team.

The centres involved — at Vanderbilt University in Nashville, Yale University in New Haven, Connecticut, and Brigham and Women's Hospital in Boston — belong to a biotechnology consortium funded by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) in Bethesda, Maryland. The source of the errors has yet to be pinpointed, and some may have arisen at the centres concerned. Troy Moore of Research Genetics maintains that the company's error rate should not exceed 2%, but adds: "If we identify a problem, we will work to correct it."

Shawn Levy, who works at the Vanderbilt centre, does not believe the problems he has found within the Research Genetics clone set are the result of local mishandling, as his team has analysed other clone sets and found error rates of less than 5%. But some of the errors might reflect the fact that cDNA clones are usually not sequenced in their entirety — so if the fragments sequenced by the NIDDK consortium do not overlap with the partial sequences deposited in public databases, correct clones may appear to be in error.

While the consortium members compare their sequencing data in an effort to pin down the source of the apparent errors, the Yale



**Teething troubles:** Timothy Zacharewski (above) was shocked by the high error rates he found in a set of cDNA clones used to make microarrays. Joseph DeRisi (right) is worried that researchers cannot check for mistakes in commercial chips.

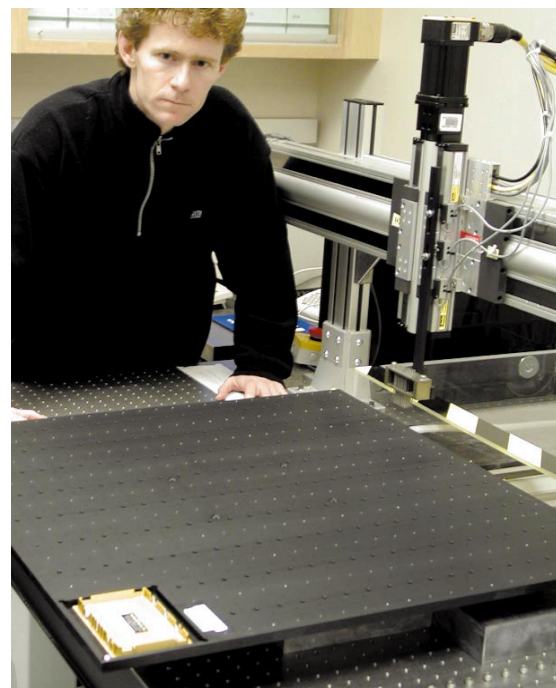
centre has posted a notice on its website warning users of the potential for problems. But regardless of the explanation, the lesson is clear: even when care is taken to remove erroneous sequences, it cannot be assumed that microarrays based on cDNA clones are reliable. "I think errors may be inherent to the system," says Steve Gullans, who heads the centre at the Brigham and Women's Hospital.

As a result, the NIDDK consortium plans to increase its output of microarrays based on a rival technology. In these chips, the grid consists of oligonucleotides, or oligos — short, single-stranded DNA segments built to order by chemical synthesis<sup>3</sup>. This construction process avoids problems with bacterial contamination, and should mean that each sequence is what the researcher orders. On the minus side, oligo-based microarrays are expensive. And ultimately, they are only as good as the information used to direct the oligos' synthesis — as the DNA chip company Affymetrix of Santa Clara, California, recently discovered.

### Mistaken identity

Affymetrix can pack up to 400,000 different oligos on a single array — usually representing around 10,000 genes, with 40 oligos for each gene. But in February, Affymetrix announced that up to a third of the sequences on one set of mouse arrays were wrong. The company had used sequences from the public sequence databases that were known to be ambiguous, and which actually corresponded to the wrong strand from the DNA double helix. As a result, the oligos could not detect their target mRNAs.

Affymetrix has promised to replace the arrays. "It's going to be an inconvenience, at most," says Carolee Barlow of the Salk Institute for Biological Studies in La Jolla, Califor-



nia, who is using the chips to investigate the genetics of brain disorders. But to DeRisi, the incident points out the risks inherent in commercial DNA chips. "You are at the mercy of the company," he says. "That is a tough situation when you are not allowed to proofread what they have done."

But even perfect arrays do not guarantee good science. Microarray experts say that some new users seem to be so mesmerized by the technology's power that they are forgetting basic principles of experimental design. Ash Alizadeh, a graduate student in Brown's Stanford lab, says he knows of several microarray studies lacking the proper controls and replications needed to ensure that differences in gene expression really are associated with the variable under investigation.

Although such shortcomings should be spotted by journal editors and reviewers, erroneous results caused by faulty chips are harder to detect — and experts are sure that some have entered the literature. They are urging users not to draw firm conclusions about the activity of individual genes without checking the sequence of the spot concerned and verifying the result using alternative methods of monitoring gene expression.

Within a few months, predicts Gullans, journal reviewers will routinely be asking these questions. And then perhaps the focus will be back on the immense power of microarrays, rather than their limitations. ■

**Jonathan Knight writes for *Nature* from San Francisco.**

1. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. *Science* **270**, 467–470 (1995).
2. Halgren, R. G., Fielden, M. R., Fong, C. J. & Zacharewski, T. R. *Nucleic Acids Res.* **29**, 582–588 (2001).
3. Lipshutz, R. J., Fodor, S. P. A., Gingeras, T. R. & Lockhart, D. J. *Nature Genet.* **21**, Suppl., 20–24 (1999).

*Nature Genetics' Chipping Forecast*

♦ [http://www.nature.com/ng/web\\_specials](http://www.nature.com/ng/web_specials)



# The story of O

Geochemists are having a hard time working out why the atmosphere of the early Earth appears to have lacked oxygen for so long. Jon Copley considers the competing theories.

**T**ake a deep breath. A fifth of the air that fills your lungs is oxygen. But it was not always like this. Around 3.5 billion years ago, the Earth's atmosphere contained almost no oxygen. Simple microorganisms had evolved, but they were adapted for an atmosphere rich in nitrogen, carbon dioxide and the sulphurous gases poured forth by volcanoes.

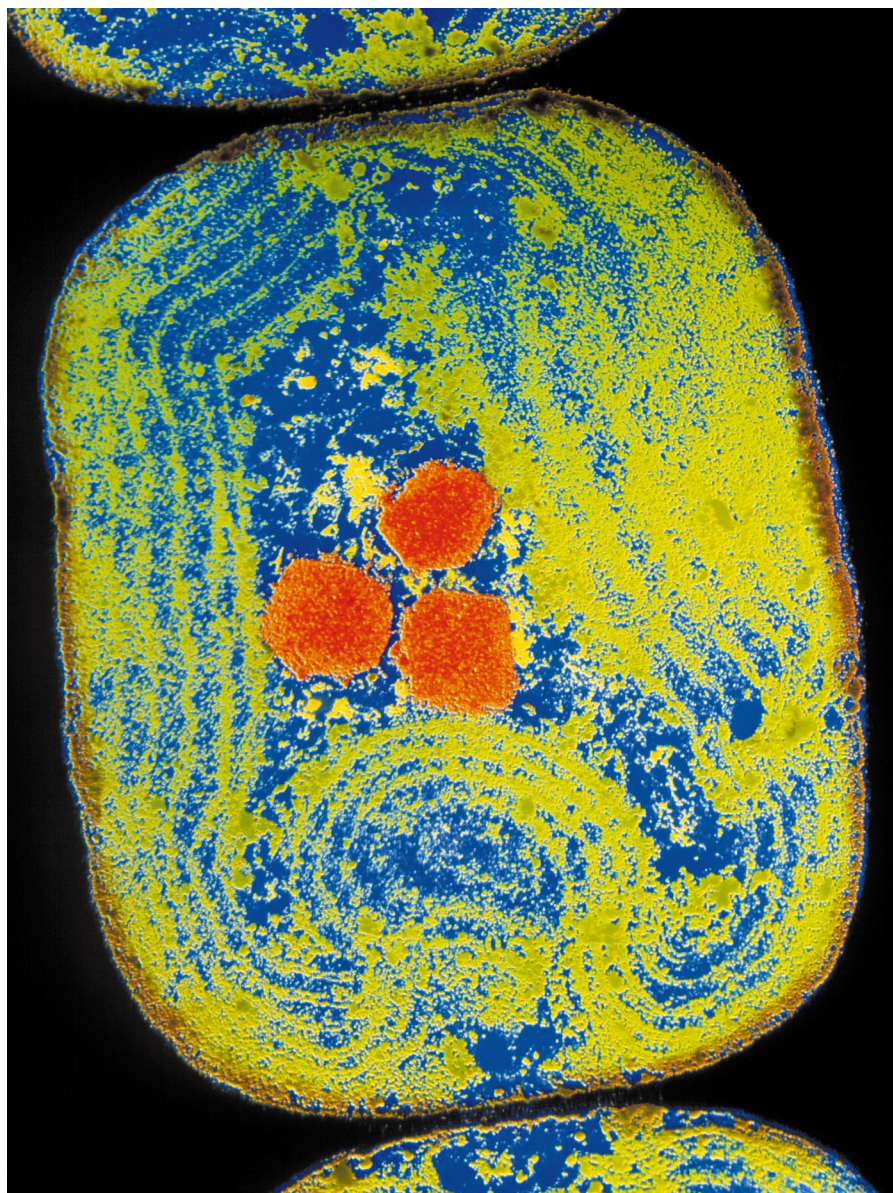
Thankfully for us, cyanobacteria evolved. These primitive microorganisms, descendants of which survive to this day, were probably the first to photosynthesize, harnessing light from the Sun to power their own growth and generating oxygen as a by-product. As a result, they started a momentous evolutionary change — adding oxygen to the atmosphere and paving the way for the eventual evolution of multicellular life.

## Up in the air

But as geologists and geochemists studying the early Earth know, the story is not quite so simple. Fossil evidence for cyanobacteria can be seen in rocks dating from as far back as 3.5 billion years ago<sup>1</sup>. But there is a host of evidence suggesting that oxygen remained a trace element in the atmosphere until about 2.5 billion years ago. So if cyanobacteria were pumping out oxygen for at least 1 billion years before it reached detectable levels, what kept the concentration of oxygen so low?

This is a difficult question to answer because the events happened so long ago. Oxygen leaves its fingerprints almost everywhere, changing the composition of rocks that are exposed to it and altering the chemistry of the oceans. It is also intimately linked to cycles of other elements such as carbon and phosphorus. But unravelling this complex web of interactions to peer back through time has taxed the ingenuity of geochemists, and the techniques at their disposal.

Researchers are helped in their search by the fact that many biological reactions occur



Early output: cyanobacteria were probably the first organisms to pump oxygen into the atmosphere.

at different rates for different isotopes of the same element. Take photosynthesis, which combines carbon dioxide and water to produce organic matter and oxygen. Almost all carbon on Earth exists as one of two isotopes — carbon-12 and carbon-13. For photosynthesis, organisms prefer to use carbon dioxide that contains the lighter carbon-12 isotope, because this means the reactions then need less energy. This preference leaves its mark billions of years later: rocks formed from organic matter contain high levels of carbon-12, whereas those formed from inorganic material include more carbon-13.

This difference provides geologists with the earliest evidence for any kind of life. Rocks from Greenland dating from 3.8 billion years ago show suspiciously high levels of carbon-13, hinting that microbes which selectively extracted carbon-12 may have been at work<sup>2-4</sup>. And Western Australian rocks formed 300 million years later provide the first fossil evidence of cyanobacteria<sup>1</sup>.

But these dates do not tally with the evidence for the rise in atmospheric levels of oxygen. Much of this evidence comes from palaeosols — rocks formed by the compres-

sion of ancient soils. The air and water of the early Earth would have left its mark on these soils and the palaeosols that formed from them, so geochemists use these rocks to test whether or not oxygen was present billions of years ago. The key to tracking oxygen is the way it reacts with iron. If there was little oxygen on the early Earth, minerals known as iron silicates would have dissolved in any water that was present, and washed straight through the soil. But oxygen converts iron silicates into insoluble iron hydroxides, which would have been trapped in the ancient soils.

## Gas detectors

Iron hydroxides have so far only been found in palaeosols younger than 2.3 billion years old<sup>5</sup>, indicating that before this time, the atmosphere did not contain much oxygen. Other mineral indicators in palaeosols seem to tell a similar story. A 2.5 billion-year-old Canadian palaeosol has been shown to contain the element cerium<sup>6</sup>. Had oxygen been present it would have reacted with the cerium to produce cerium oxide — but none of this oxide was found.

How can this apparent paradox between





**Puzzling evidence:** the formation of pyrite (above) and banded iron formations (left) indicate that atmospheric oxygen levels did not rise until at least 1 billion years after photosynthesizing cyanobacteria had evolved.

NATURAL HISTORY MUSEUM, LONDON

the emergence of cyanobacteria and the rise in oxygen levels be explained? Some researchers argue that the paradox does not exist and that the interpretation of the evidence is wrong. Others invoke changes in the way that organic material released by dead microorganisms reacted with oxygen. And a new hypothesis suggests that changes in gases emitted by volcanoes could be the key.

Hiroshi Ohmoto, a geochemist from Pennsylvania State University, is one of those who thinks that there is no paradox, denying that oxygen emerged as late as the palaeosol evidence suggests that it did. Ohmoto argues that the lack of iron in palaeosols does not mean it was never there. It might have been removed by hot water from volcanoes<sup>7</sup>.

Organic acids produced by bacteria might also have stripped iron from palaeosols. Ohmoto says he has seen a number of palaeosols where this might have happened. The organic acids may have been produced by cyanobacteria. Although cyanobacteria first evolved in the oceans, they could soon have colonized the land. Ohmoto says that a 2.6 billion-year-old South African palaeosol looks as if it may have had a large community of cyanobacteria living in it<sup>8</sup>. So these microorganisms could be confusing the geological record — creating the oxygen that fixed iron in soils and also producing the acids that removed it. Put these factors together, says Ohmoto, and the timing of the jump to high oxygen levels is “very, very difficult to clarify”.

Ohmoto is currently working on a geochemical model that appears to back up his assertions. In unpublished studies conducted with Antonio Lasaga, a geochemist until recently at Yale University in New Haven, Connecticut, Ohmoto has modelled the cycles of oxygen, carbon, sulphur and iron for the early Earth. “The end result is very striking,” he says. “Within 30 million years from the time cyanobacteria evolved, the atmosphere will contain very high oxygen content and will remain basically the same.”

Thirty million years is the blink of a geological eye, so if cyanobacteria appeared at 3.5 billion years ago or earlier, the atmosphere would have contained large amounts of oxygen from then onwards.

### Rocky road to success?

But if Ohmoto is going to convince other researchers that he is right, he will have to account for a range of other evidence that points to low oxygen levels before 2.5 billion years ago. At that time, minerals such as pyrite and uraninite were being washed around by rivers. Contact with oxygen would have altered their composition, but sediments from 2.75 billion years ago and earlier contain the minerals in their unreacted form<sup>9</sup>.

More evidence comes from deposits that cannot form if oxygen is present. Examples of one of these — banded iron formations — seem to be confined to rocks dating from more than 2.3 billion years ago, indicating that oxygen was not present at that time. But Ohmoto points to a similar banded iron formation dating from 1.8 billion years ago as evidence that such structures can form in the presence of oxygen. Other researchers argue that this formation is probably an anomaly, arising in water that contained little oxygen.

A new approach to the problem has further strengthened the case for the late emergence of oxygen. Developed by James Farquhar of the University of California, San Diego, the technique takes advantage of a quirk in the behaviour of different isotopes. Biological reactions such as photosynthesis leave clear signatures in rocks by selecting for lighter isotopes. Now it appears that some non-biological reactions also select for certain isotopes, although here, mass is not the governing factor. Although the selection mechanism — known as mass-independent fractionation — is not fully understood, the effects of it can be seen in the geological record.

Sulphur gases in the Earth's early atmosphere were subject to both biological and non-biological fractionation. Microorganisms that fed on sulphur preferred the lighter of the element's three isotopes and the resulting rocks reflect this. But some of the non-biological processes involved in the formation of other rocks can cause mass-independent fractionation of sulphur. Such rocks contain isotopes of sulphur in distinctive ratios that do not fit with those generated by biological processes.

Crucially, this mass-independent selection of sulphur is driven by ultraviolet (UV) light from the Sun. As oxygen appeared in the atmosphere it would have slowed the reaction down by forming a layer of ozone — a molecule containing three oxygen atoms — which blocked out much of incoming UV light. Oxygen would also have combined with sulphur and removed it from the atmosphere. Together, these effects would have stopped the mass-independent selection of sulphur. And according to Farquhar, evidence for mass-independent fractionation in the geological record ceases at 2.45 billion years ago<sup>10</sup>.

For Jim Kasting, a geochemist at Pennsylvania State University, Farquhar's evidence is “the real clincher”. Heinrich Holland, a geologist at Harvard University, agrees and says he will argue so when he reviews the evidence at a joint meeting of the Geological Societies of London and America to be held this summer in Edinburgh. “Every time somebody looks at something different, it's one more piece of evidence for low oxygen before 2.3 billion years ago,” he says.

But Ohmoto is not willing to give up the fight just yet. He says that recent work<sup>11</sup> from Don Canfield of the Danish Center for Earth System Science in Odense backs his early oxygen theory. Canfield has found evidence for sulphate-devouring microorganisms at around 3.5 billion years ago. Oxygen in the atmosphere promotes the production of sulphate in the oceans, leading Ohmoto to argue



## news feature

that evidence of the microbes is evidence of oxygen. But other researchers, including Canfield himself, are not convinced that one implies the other, reasoning that other mechanisms could have produced the sulphate.

### Into the black

But if the evidence does favour the late emergence of oxygen — and most geochemists believe it does — the original question remains unanswered: why did oxygen take so long to build up after cyanobacteria first emerged? The answer may lie not in the processes that create oxygen, but in the ones that mop it up. “It’s like continually having your spending exceed your income — it does not lead to wealth,” explains Holland.

The organic carbon left over after the decay of dead microbes can react with oxygen. The critical factor is how much of this organic matter gets buried underground, away from oxygen sources in the oceans and atmosphere. Rocks formed around 2.2 billion years ago show a peak in carbon-12 levels, indicating that the burial of organic carbon increased at this time<sup>12</sup>. Researchers believe that movement of tectonic plates may have opened up deep basins on the ocean floor that were free of oxygen, where organic material could have settled<sup>13</sup>.

Volcanic gases could also have mopped up oxygen. According to new work from Kasting and his colleagues<sup>14</sup>, a change in the composition of these gases about 2.7 billion years ago may have had a dramatic effect on oxygen concentrations.

Kasting’s model depends on the compo-



**Smokin’:** changes in volcanic emissions could have tipped the balance and allowed oxygen levels to rise.

sition of the Earth’s mantle — the region below the crust and above the core, where volcanic gases originate. Before the emergence of oxygen, the mantle was rich in ‘reducing’ minerals such as iron silicates, which react with oxygen when exposed to it. A reducing mantle would have produced reducing gases which, when released into the atmosphere, extracted much of the oxygen produced by photosynthesis.

### Late reductions

But the mantle is not static. At its bottom, heat from the Earth’s core forces material upwards, whereas at the Earth’s surface the crust and the mantle immediately below it slide back underground. Material that breaks the surface comes into contact with whatever oxygen is present in the atmosphere and oceans, reacts with it and loses some of its reducing strength. So over time, this ‘oxidized’ material would move back down to the base of the mantle where it would build up. According to Kasting, the mantle underwent a major switch around 2.7 billion years ago pushing the non-reducing material to the surface. As a result, the volcanic gases emitted by the mantle would have become less reducing. Kasting and his colleagues calculate that by 2.7 billion years ago these gases had lost much of their ability to mop up atmospheric oxygen, allowing its levels to rise.

Holland is currently working on a similar issue. He is looking at the contribution to volcanic gases arising from downward moving rocks in what are known as subduction zones. Subduction material emits gases that are less reducing than those coming from the mantle into which it sinks, and Holland believes that

the contribution to volcanic gases from subduction zones was growing at just the time that oxygen levels started to increase.

The new work by Holland and Kasting may yet join more established theories such as carbon burial. Put together, they could explain how the atmospheric oxygen budget broke from the red into the black. The issue might not be settled yet, but the geological record is starting to make more sense.

But for life on Earth, the emergence of oxygen was only the beginning of the story. Oxygen’s first effect was to make conditions harsher — combining with and removing much of the methane that had trapped heat in the atmosphere. Within a few million years, ice and snow covered the Earth, killing much of the primitive life that had evolved. Indeed, the emergence of oxygen was just the first step in the dance between life, geology and climate that continues to this day. ■

**Jon Copley is a freelance writer, and a teaching fellow at the University of Southampton.**



**Digging in:** Hiroshi Ohmoto believes that oxygen levels rose around 3.5 billion years ago.

1. Schopf, J. W. *Science* **260**, 640–646 (1993).
2. Schidlowski, M., Appel, P. W. U., Eichmann, R. & Junge, C. E. *Geochim. Cosmochim. Acta* **43**, 189–199 (1979).
3. Mojzsis, S. J. et al. *Nature* **384**, 55–59 (1996).
4. Rosing, M. T. *Science* **283**, 674–676 (1999).
5. Holland, H. & Rye, R. *Geology* **25**, 857–859 (1997).
6. Murakami, T., Utsunomiya, S., Imazu, Y. & Prasad, N. *Earth Planet. Sci. Lett.* **184**, 523–528 (2001).
7. Ohmoto, H. *Geology* **24**, 1135–1138 (1996).
8. Watanabe, Y., Martini, J. E. & Ohmoto, H. *Nature* **408**, 574–578 (2000).
9. Rasmussen, B. & Buick, R. *Geology* **27**, 115–118 (1999).
10. Farquhar, J., Bao, H. & Thiemens, M. *Science* **289**, 756–758 (2000).
11. Shen, Y., Buick, R. & Canfield, D. E. *Nature* **410**, 77–81 (2001).
12. Karhu, J. A. & Holland, H. D. *Geology* **24**, 867–870 (1996).
13. Des Marais, D. J., Strauss, H., Summons, R. E. & Hayes, J. M. *Nature* **359**, 605–609 (1992).
14. Kump, L. R., Kasting, J. F. & Barley, M. E. *Geochim. Geophys. Geosyst.* **2** (2001). <http://146.201.254.53/publications/final/researchletters/2000GC000114/fs2000GC000114.html>

## Biotech offers Africans a chance to create their own practical solutions

*Sir*—Agricultural biotechnology research and development (R&D) in industrialized countries is heavily supported by private and government institutions and universities, which develop products and services for capital-intensive farming systems. Although some innovations have spill-over effects that might benefit Africa (especially large-scale farmers), most are likely to marginalize poor farmers.

One solution is to reorient international biotechnology research to take account of small-scale farmers' needs. But experience shows that this option is unlikely to succeed in the long term.

A more realistic way forward is for African scientists, businesses and farmers to devise biotechnological innovations that are appropriate to local cultural, economic, political, technological, institutional, infrastructural and social factors.

The success story of Africa Online, the premier provider of Internet services throughout Africa, illustrates an entrepreneurial spirit that should be possible to replicate in the biotechnology arena. Africa Online was founded in 1994 by three Kenyans studying at the Massachusetts Institute of Technology and Harvard University. It now serves thousands of people and businesses throughout Africa.

Private-sector companies wishing to invest in biotechnology are attracted by successes in tissue-culture-aided production and multiplication of disease-free planting materials for cassava, yam, banana, plantain, citrus and flowers in countries such as Kenya and Ghana. However, Africans must learn simple technologies that are not only appropriate and feasible, but also sustainable.

Priority can be given to biotechnologies that have worked under comparable conditions elsewhere. Last year, for example, thousands of poor Chinese farmers obtained up to 40% increases in sweet-potato yields by using a novel seed-production technique to eliminate viral diseases from planting materials ([http://www.futureharvest.org/growth/china\\_sweet.bkgnd.shtml](http://www.futureharvest.org/growth/china_sweet.bkgnd.shtml)). No genetic improvements were made, and the farmers used no more fertilizers or pesticides than usual.

In Africa, the key players will include, among others, scientists, policy-makers, non-governmental organizations, farmers and farmers' (especially women's) groups. Private biotechnology initiatives must go hand in hand with development of regulatory frameworks and public-

awareness campaigns, and with other R&D programmes targeting poor farmers (see, for example, F. Wambugu, *Nature* **400**, 15–16; 2001)).

The mistake of the Green Revolution was that it treated all the world as if it were the same. The lesson for agriculture is that problems must be solved locally and communally, through a bottom-up approach that empowers farmers to support and own technologies that benefit them. Entrepreneurial scientists, business-people, lawyers and farmers are needed to explore the promises of biotechnology.

This is a wake-up call for African biotechnology stakeholders to transform the potential wealth of genetic resources and traditional knowledge into the reality of increased incomes — and better food and health care — for the majority.

**Jesse Machuka**

*Biotechnology Laboratory, International Institute of Tropical Agriculture, PMB 5320, Ibadan, Nigeria*

## Kansas science saved by teachers' good sense

*Sir*—The News in Brief report "Evolution allowed back into Kansas schools" (*Nature* **409**, 973; 2001) is not entirely accurate. In 1999, the Kansas Board of Education approved ill-advised science standards that did not include several key scientific principles, including macroevolution and the Big Bang.

However, contrary to many reports in the popular press over the past two years, the board does not have the power to direct schools to teach anything. The function of the standards is only to set the guidelines used in state-wide standardized testing, and they are used by teachers as a guide to help prepare their curricula.

Thanks to the good sense of Kansas teachers, the teaching of evolution was not left out of the curriculum in schools during the year these unfortunate and controversial standards were in effect. In fact, some anecdotal evidence suggests that evolution has been emphasized more than ever in Kansas schools since then.

The assertion in your News report that "creationists ... briefly had the Bible's account of the beginning of the world taught instead" is untrue. There was never any mention of the biblical account of creation in the controversial standards, and there have been no reports that creation was taught as part of any public school science curriculum in Kansas.

The News report correctly states that the citizens of Kansas have rectified the situation by defeating several of the anti-evolution board members in last year's elections, and that new science standards

are now in place that include the aforementioned key scientific principles.

**Rollie J. Clem**

*Division of Biology, 307 Ackert Hall, Kansas State University, Manhattan, Kansas 66506, USA*

## Bright light of learning snuffed out in Breslau

*Sir*—Vaclav Smil's delightful Millennium Essay "*Genius loci*" (*Nature* **409**, 21; 2001) emphasizes the importance of Budapest in twentieth-century science. I would like to propose Breslau, Germany (now Wrocław, Poland), as another locus in central Europe. All the following scientists were born or grew up in Breslau: Max Born (Nobel Prize in Physics, 1954), Richard Courant (mathematician, 1888–1972), Paul Ehrlich (Nobel Prize in Physiology or Medicine, 1908), Fritz Haber (Nobel Prize in Chemistry, 1918), Reinhard Selten (Nobel Prize in Economic Sciences, 1994), Otto Stern (Nobel Prize in Physics, 1943) and Otto Toeplitz (mathematician, 1881–1940).

The scientific glory of the city is now lost. All in the above list were of Jewish descent and most were forced to leave Germany during the Nazi period. During and after the Second World War, the entire composition of the city changed.

Reinhard Selten in his Nobel prize autobiographical sketch (see <http://www.nobel.se/economics/laureates/1994/selten-autobio.html>) says, "I have never visited Wrocław since the war. Heavy fighting destroyed most of the town in which I grew up and most of the familiar places of my youth look different now."

**Min-Liang Wong**

*Department of Veterinary Medicine, College of Veterinary Medicine, National Chung-Hsing University, Taichung 402, Taiwan*

## New information on the biodiversity facility

*Sir*—Your otherwise interesting News report (*Nature* **410**, 290; 2001) on the new Global Biodiversity Information Facility (GBIF — see <http://www.gbif.org>) contains some errors. GBIF, which came into being on 1 March this year, was not launched in Brussels. The first meeting of the GBIF governing board took place on 9–11 March in Montreal, Canada. Martin Sharman and Carlos Martinez-Riera are scientific officers at the European Commission; they do not work for GBIF.

**Christoph L. Häuser**

*(Chair, GBIF Governing Board)  
State Museum of Natural History,  
Rosenstein 1, D-70191 Stuttgart, Germany*



# A year of opportunity

Now is the time for the UK Food Standards Agency to flex its muscles.

## Eileen Rubery

How can hygienic, safe, food-animal husbandry most effectively be combined with modern farming practices? Widespread public concern over genetically modified foods, bovine spongiform encephalopathy (BSE) and other safety issues are being thrown even more sharply into relief by the present foot-and-mouth disease epidemic which started in the United Kingdom, but is having significant repercussions for the rest of Europe. What is it about the United Kingdom that has given rise to the recent spate of food-related problems? One year ago this month, the Food Standards Agency (FSA) was created to fulfil the manifesto commitment of the Labour government, with a remit to act as an independent voice for the consumer. Now that it has had its first 'success' — its review of BSE controls (see Boxes 1–4) — and its settling-in period is over, it is time to ask how the agency is going to attack its remit — to make a real difference to food safety.

I suggest two priorities for the agency's second year. First, that it builds on the ground-breaking (for the United Kingdom, at least) approach it used for its BSE controls review, and ensures that all its activities are carried out with at least equal transparency and openness; and second, that it now identifies a 'big issue' to tackle that will really make a difference to food safety. Given the present foot-and-mouth disease epidemic, a review of on-farm, and in particular, food-animal practices, to determine how the risk of infection and contamination of farm products with potential human and animal pathogens can be minimized from 'birth (or planting) to plate' is a compelling candidate.

## Changing food management

The foot-and-mouth disease epidemic (see *Nature* 410, 501 and 515; 2001) is drawing the world's attention to the fact that moving animals around as part of 'growing' them is potentially a highly efficient way of spreading animal or human pathogens across Europe. Undoubtedly, once the UK epidemic is under control, there will be a move to reassess farming and food-management practices. This reassessment should be sponsored by the independent, consumer-focused FSA, not by the Ministry of Agriculture, Fisheries and Food (MAFF), in recognition that the purpose of the process is to produce safe, high-quality food. Any such review should be conducted with the same degree of openness as the BSE controls review (see Boxes 1 and 2), but with greater publicity for the open meet-

## Box 1 The FSA review of BSE controls

- **First major task of the FSA**, at personal request of the UK Prime Minister, Tony Blair.
- **Terms of reference** — "to review the current main measures to protect the public against BSE/vCJD in relation to the food chain..." — a challenging, politically sensitive task.
- **Review team chaired by the chairman of agency**, augmented by external advisers, including two from mainland Europe.
- **First action** was to set up a **stakeholder group** of representatives of the relevant industries, the academic community, consumer groups and a range of other key bodies (see <http://www.bsereview.org.uk>), which **met in public** to discuss the issues and advise the review team.
- The **'public'** included representatives of vCJD patient groups, those affected by the BSE epidemic, consumer groups and academics. There was also a website for comments.
- **The public drew attention to important issues** including the composition of 'natural' sausage skins, imported foods and risk assessment, resulting in a stronger final report.

ings and greater efforts to get participation from all relevant groups.

By acting swiftly now and setting its review in place, the United Kingdom could occupy a key position in the inevitable European Community review into the way we manage farms in the production of high-quality, safe food. The Agence Française de Sécurité Sanitaire des Aliments, France's food-safety agency (<http://www.afssa.fr>), has pursued a strongly independent line on whether beef poses a health risk; a consulta-



Making a difference: workers for the FSA take food samples away for testing.

tion paper proposing a European Food Authority was published by the European Commission on 8 November, but adoption is unlikely before the end of this year.

Getting the recommendations from such a review right could deliver great benefits for farmers, consumers and the food industry throughout Europe. The FSA is well placed to lead this initiative, which might do much to rehabilitate Britain among its European neighbours from having been the conduit, however unwittingly, of the foot-and-mouth

## Box 2 Unique features of the BSE report

- **Sheep** Discussions on whether sheep are infected with BSE (as opposed to scrapie) were held in the open. The first draft stated that the data needed to decide whether some sheep had BSE were not available, concluding that if such evidence did appear, then specified risk material (SRM) controls similar to those for cattle would not be appropriate because prions are much more widely distributed in sheep tissue than in cattle.
- **Being up-front about uncertainty** The difficulties in predicting the size of the vCJD epidemic in humans were clearly stated.
- **Openness** The FSA maintained its policy of openness in the face of considerable, potentially hostile, media interest. Each stakeholder discussion and revised draft evoked new media interest. Industry members of the stakeholder group were initially concerned, but consumer confidence in lamb held.
- **Imported meat** The difficulties of controlling imported meat products and ensuring they did not include beef from cattle over 30 months old were reported in the media. By the final report, Germany and other mainland European countries were also discussing the issue.
- **Views from the website** One annex summarized views from the website and other sources of comment.

virus epidemic, and the origin of BSE. Both these epidemics have had, and will continue to have, serious global consequences.

Deciding what changes are needed is far from easy. Elimination or reduction in human and other animal pathogens requires well-thought-out, practical programmes, enforced sympathetically but firmly, and implemented by committed farmers. The farming community and the food industry must work together to develop a holistic approach to food production. Right from the first, risk assessments and cost-benefit analyses need to be considered from the viewpoint of their effect on food safety and animal health at all points in the food chain, not just from a local angle.

### Controlling contamination

The United Kingdom has for too long taken a pessimistic view of what can be achieved with respect to on-farm hygiene. Whatever the pathogens carried by the live animal, we have generally assumed it is the responsibility of the food processor to get rid of them — for example, in the case of raw chickens, the domestic consumer has been expected to kill any pathogens by cooking.

There have been advances: since the 1980s the UK poultry industry has made real efforts to reduce *Salmonella* contamination of chickens. Some retailers now believe these efforts are paying dividends. In contrast, cows, pigs and sheep are still largely farmed without any serious attempt to prevent the spread of human pathogens between them. The United Kingdom's recent *Escherichia coli* O157 outbreak (before the FSA was established) is an example of the occasional consequence of infected animals going for slaughter.

The concept of controlling pathogenic contaminants of food during food processing and preparation using a 'hazard-analysis system' for assessing hazards and controlling risks is well established; and is included in some European Union legal requirements. The FSA should now examine how best to apply this philosophy to food animals and crops on the farm.

Some work may already be in hand. The

### Box 3 Consequences of the openness policy

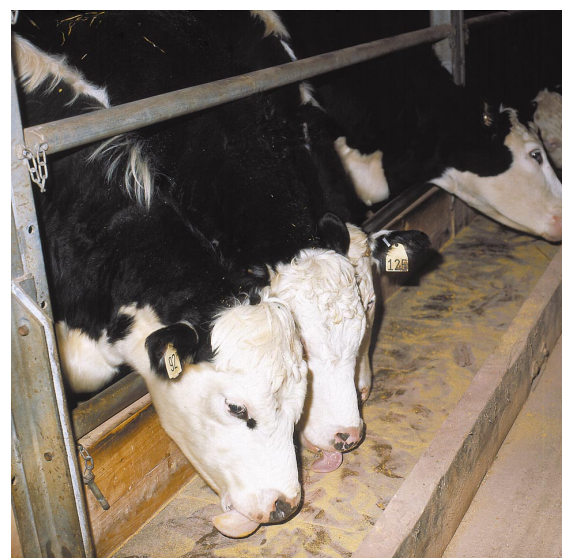
- Media coverage of the drafts was high. Publication of the BSE report itself was rather a non-event, as it contained little not already in the public domain.
- Could this lack of media interest encourage government ministers to feel let off the hook in implementing recommendations?
- There have been more than a million hits on the FSA website since September 2000; 5,000 copies of the final report have been downloaded.
- Anyone can register for further information on progress of the review ([www.bsereview.org.uk](http://www.bsereview.org.uk)).
- The website reports several infractions of the BSE controls detected in beef imported from mainland Europe, so one action recommended in the report is being implemented.

UK Chief Veterinary Officer, Jim Scudamore, set up a zoonoses committee to investigate this area in 1999. The work of this committee needs to be given a higher profile. How can a culture of hazard analysis and control for human as well as other animal pathogens on-farm improve food safety? Similarly, crop management should be assessed in the context of the spreading of manure and other potentially contaminated material.

The FSA should reconsider the conclusions of the report commissioned by the 'shadow' FSA (before the agency was a legal entity) that there is no need for an 'on-farm' focus in the agency. Consumer groups fought hard for the FSA to have an on-farm remit.

### Transparency and the FSA

The openness of the FSA's BSE review was refreshing and stimulating (see Boxes 1 and 2). For Britain, such a procedure for considering an issue and developing advice was almost revolutionary, and an important change from the past. As evidence of a will for openness and wide consultation, the entire experience appeared to me, a member of the review team, very positive. But it can only be the first step



Food chains: animal husbandry has come under increasing scrutiny since the BSE outbreak.

towards a robust culture of openness.

No culture can change overnight. Officials involved in routine reviews will naturally tend to continue working in the same way, unless there is a clear, sustained mandate for a new culture. Other ministerial departments may not support this stance if it affects their policies. Constant encouragement and support is needed from the FSA board and senior members of the executive (see Box 4) to ensure that the commitment to openness eventually takes root throughout the entire organization.

The US Food and Drug Administration publishes regular notices of the issues it intends to review, and open hearings are the rule rather than the exception. The US Freedom of Information Act ensures ready access to the essential technical data, enabling pressure groups to seek informed advice from experts so that the quality of their interventions can be high. It remains to be seen if Britain's long-promised legislation will provide the power to achieve similar results.

My verdict on the FSA after its first year is that it has made a promising start. But year two is crucial. The problems of foot-and-mouth disease, concern about *Mycobacterium paratuberculosis* in milk, *E. coli* O157, other microbial contamination and BSE are stark evidence of the need to focus on all stages of the food chain, including farming practices. The agency must now use its power to really make a difference there as well as elsewhere.

Eileen Rubery is at Cambridge University Health, Judge Institute of Management Studies, University of Cambridge Business School, Trumpington Street, Cambridge CB2 1AG, UK. She headed the protection of health division in the public health group of the Department of Health until July 1999. She was an independent expert member of the FSA review team for the BSE controls review. She is an expert adviser to the FSA and to other UK government departments from time to time.

### Box 4 How 'independent' is the FSA?

- The agency is a non-ministerial department (<http://www.foodstandards.org.uk>) reporting to parliament through health ministers.
- The chair of its board is Sir John Krebs (formerly head of the UK Natural Environment Research Council and a professor of zoology at Oxford University).
- The executive is about 600 civil servants, originally recruited from the Department of Health and MAFF and headed by Geoffrey Podger, plus the Meat Hygiene Service, responsible among other things for the inspection of abattoirs.
- Between Sir John and the executive sits the independent board.
- Renaming and relocating a group of people will not change their behaviour. The board and executive have set a style in their first review, but changing cultures takes time and effort.
- The FSA website reports an open meeting on genetically modified foods last month; some advisory committees hold open meetings.



# How persistence paid off

The struggle, delusion and arrogance of the character who found Java Man.

## The Man Who Found the Missing Link: Eugene Dubois and His Lifelong Quest to Prove Darwin Right/The Extraordinary Life of Eugene Dubois

by Pat Shipman

Simon & Schuster/Weidenfeld & Nicolson:  
2001. 514 pp. \$28/£25

Douglas Palmer

In her latest book, science writer and anthropologist Pat Shipman gives us a blow-by-blow account of the life of Eugène Dubois, the Dutch physician and anatomist who in 1892 discovered 'Java Man', *Pithecanthropus erectus*, now known as *Homo erectus*. Shipman's account of Dubois' story is fascinating; much of it is set in an exotic location, and it is important for the insights it gives into the development of anthropology and the study of human evolution.

Dubois' long life started in 1858, the year before the publication of Charles Darwin's *On the Origin of Species* and five years before the first formal recognition of an extinct human species, *Homo neanderthalensis*, by William King in 1863. Dubois died in December 1940, the year German troops occupied the Netherlands. He was still fighting his corner, protesting about the paper by Ralph von Koenigswald and Franz Weidenreich in *Nature* (144, 926–929; 1939), which suggested that *Pithecanthropus* and *Sinanthropus* were two representatives of the same prehuman stage.

The story of why Dubois gave up a promising academic career at the University of Amsterdam, and dragged his family all the way to the Dutch East Indies in search of the so-called 'missing link' certainly has its own attraction. He had been inspired by Alfred Wallace's writings on Java's indigenous animals and plants and by Ernst Haeckel's theory — based on the belief that the Indonesian orang-utan was most closely related to humans — that humankind originated in Southeast Asia. After several fruitless years, Dubois found what he was looking for — the result of astonishing luck and sheer persistence. The years of struggle to get 'his' find acknowledged and recognized by other scientists is also an extraordinary story of struggle, bitter disappointment and delusion, combined with peculiar arrogance.

The ultimate pathos and tragedy of Dubois' story is painfully well portrayed by Shipman. By the time his find was accepted as a valid taxon, the world had moved on and Dubois' Java Man had been overshadowed by the newer and more spectacular finds of *Sinanthropus*, or 'Peking Man', now also



Object of derision: Dubois' drawing of the skull of his *Pithecanthropus erectus*.

regarded as *Homo erectus*. Nevertheless, as Shipman demonstrates, Dubois made significant contributions in other areas of anthropology, especially in his studies of cephalization — the evolution of the head and brain — and the morphometrics of brain and body size.

Shipman is at her most successful when quoting from Dubois' surviving letters and journals. There is, however, a certain amount of reconstructed dialogue and inner monologue of the following sort: "There, Dubois thinks as he pens the last lines, I have said it. I have stated my credo, my opinions, and it remains for the world to judge who is right. But it is I who will be remembered as the man who found the missing link." I found this tedious at times, even if there is good evidence for its validity, but other readers may be more tolerant.

Father J. J. A. Bernsen, a Jesuit priest, was Dubois' assistant in 1930–32, and Shipman has used his diaries, in which "many conversations with Dubois are recorded apparently verbatim". The hapless Father Bernsen spent two years cataloguing Dubois' collection of

more than 10,400 mammal bones from the Trinil site where *P. erectus* had been found and hoped for some glimmer of recognition from Dubois. Bernsen even found eight more fragments of *P. erectus* bones that had been overlooked, but he got no thanks from Dubois. Quite the reverse — his find showed that one of Dubois' key points, that all the hominid bones at Trinil came from the same individual, was wrong. Dubois bullied Bernsen and plotted to get him removed. Poor Bernsen was taken ill and died of abdominal bleeding, unthanked and unacknowledged by Dubois, who constantly complained of how poorly others treated him.

Shipman ends Dubois' story thus: "he left behind many enemies ... his story became notorious, told and mistold many times until it reached mythic proportions. It is now time for the truth, and I have told it." But although Shipman does set the record straight, and the often-prolonged traipsing through daffodil-filled fields of dialogue might well capture the spirit of the moment, can it be 'the truth'? If, however, she means the more problematic 'truth' about Dubois'

## book reviews

unpleasant persona, then I agree. It must often have been difficult to write a biography whose hero turns out to have been quite a monster, however fascinating. ■

Douglas Palmer is at 31 Mawson Road, Cambridge CB1 2DZ, UK.

## Slaves to logic

### **The Universal Computer: The Road from Leibniz to Turing**

by Martin Davis

W. W. Norton: 2000. 257 pp. \$26.95, £18.95

### **Computers Ltd: What They Really Can't Do**

by David Harel

Oxford University Press: 2000. 214 pp. £14.99, \$25

John Naughton

John Maynard Keynes once observed that "practical men ... are usually the slaves of some defunct economist", although this may be regarded as special pleading from a practitioner of the "dismal science" — as Thomas Carlyle described economics.

Nowadays everyone — or at any rate everyone who uses even a word-processor — is a slave of a defunct logician.

For computers are essentially logical engines rather than arithmetic calculators. This was Alan Turing's great insight, but it was in fact the culmination (or even a logical conclusion?) of a continuous stream of thought stretching back to Gottfried Leibniz. He not only built one of the earliest mechanical calculators, but also dreamed — according to Martin Davis — of finding "a special alphabet whose elements represented not sounds but concepts. A language based on such an alphabet should make it possible to determine by symbolic calculation which sentences written in the language were true and what logical relationships existed between them."

Davis's book is a charming unravelling of the intellectual thread that links Leibniz to Turing. The journey takes him first to the work of George Boole, who turned logic into algebra (and made one and one equal one, to the delight of generations of schoolboys). Then he surveys the despairing life of Gottlob Frege, who provided the first fully developed system of logic that encompassed

all of the deductive reasoning in ordinary mathematics, and yet had his life's work destroyed by a letter from Bertrand Russell pointing out an error in one of his fundamental assumptions.

Then there is Georg Cantor, who tried to create a coherent mathematical theory of the infinite. After him comes David Hilbert, who reduced the consistency of euclidean geometry to that of arithmetic, but left to others the problem of establishing the consistency of arithmetic.

After Hilbert, the baton passed to Kurt Gödel, who showed not only that there was a meaningful notion of mathematical truth, but also that it extends beyond what can be proved in any given formal system. Which brings us to Turing, who conceived of the all-purpose computer — the celebrated but infinitely tedious 'Turing machine' with its endless tape and relentless logic.

The task of hacking a readable path through the dense thickets of symbolic logic is not for the faint-hearted. It requires the inner confidence of the real expert — the chutzpah that enables a teacher to dispense with 95% of the detail and focus on the 5% that really matters. *The Universal Computer* reeks of that kind of confidence, and as a result succeeds admirably in conveying the broad sweep of its subject over three centuries. And it breaks with convention by dwelling not only on the intellects but also on the personalities of its heroes. Frege, for example, may have been a saint in intellectual terms, but he was clearly a monster in the flesh. Gödel was a tortured soul whose mental stability became exceedingly fragile towards the end of his life. And even Turing was not without his demons — although they may have had more to do with sexuality than mathematics. The reader is left with the feeling that while one clearly has to be very clever to make serious contributions to symbolic logic, one might have to be very peculiar to want to.

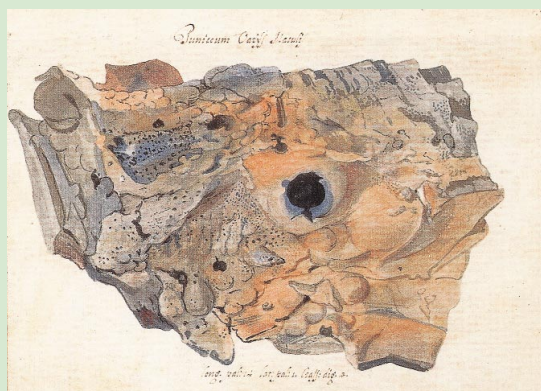
The fact that computers are logical engines, together with contemporary hype about Moore's law (which states that computing power doubles every 18 months), may have given rise to the popular fantasy that there are no limits to what one can achieve with software and the right hardware. If anyone does indeed harbour such illusions, then David Harel's book should be an excellent antidote. Its purpose is to explain why and how certain kinds of problems are not — and cannot be — solvable by computation.

Some of his material is adapted from an earlier, more technical exposition (*Algorithms: The Spirit of Computation*, Addison Wesley, 1987), which likewise expounded what Harel calls the "bad news" about computing. *Computers Ltd* is crisply written for a popular audience and provides an excellent overview of the nature of algorithms and their



## Preserved for posterity

A view of Italian fossil wood deposits (above) and a specimen of baked lignitic clay, from *Fossil Woods and Other Geological Specimens* by Andrew C. Scott and David Freedberg (Harvey Miller, £150, \$225), the latest volume in the catalogue of the seventeenth-century *Paper Museum of Cassiano dal Pozzo*.





applications and limitations. It gives a good intuitive account of what's involved in judging whether a given problem is computable, and an excellent explication of the Turing machine and the Church–Turing thesis (which says that Turing machines are capable of solving any effectively solvable algorithmic problem). It dips into parallel and quantum computing and cryptography and discusses the kinds of puzzles (the tiling problem, Towers of Hanoi, the monkey puzzle) that keep whole families fractiously occupied over Christmas. Whether it will induce an appropriate sense of humility in the computer industry, however, remains to be seen. ■

John Naughton is at the Open University, Milton Keynes MK7 6AA, UK. *A Brief History of the Future*, his book on the history and significance of the Internet, is published by Phoenix and Overlook Press in the United Kingdom and United States, respectively.

## Light on a distant subject

### Star Clusters

by Bruce W. Carney & William E. Harris  
Springer: 2001. 417 pp. £44.50, \$69.95

### Carla Cacciari

Not long after the Big Bang occurred, galaxies started to form. In this process stars also formed, as isolated individuals or bound into groups of up to several million stars. These spheroidal structures are called globular clusters and are interesting as fossil records of the early Universe. Younger clusters have also been found, showing that star formation can occur at later stages in the life of a galaxy. As all stars in a cluster are coeval and chemically alike, globular clusters are excellent laboratories for studying stellar evolution and for testing the models and physical assumptions underlying the theories. And last but not least, the properties of the cluster luminosity function are remarkably invariant from galaxy to galaxy, and so provide an excellent standard candle for determining distances.

Every year since 1971 the Swiss Society of Astrophysics and Astronomy has held a one-week advanced course on an astronomical subject of general interest. *Star Clusters* is based on the lectures on globular star clusters given by Bruce Carney and William Harris at the 1998 course.

Bruce Carney's lectures are a good introduction to the evolution of stars in globular clusters. We meet the first basic concept in astrophysics: that stars that are born bigger evolve faster than less massive ones. Thus, we see stars in different evolutionary stages that tell us what it is like to be a teenage, adult or even a dead star in, say, a 14-billion-year-old



Globular clusters such as G1, in the Andromeda galaxy, tell us much about the early Universe.

globular cluster. Carney leads us skilfully through various technical details, unusual objects, and physical processes from basic nucleosynthesis to specific mixing phenomena in the stellar interiors. Along the way he discusses chemical and dynamical characteristics and their implications, which lead to the key item, the ages of the clusters. Relative ages among globular clusters, along with their chemical and dynamical properties, show us how the Galaxy formed, whereas the clusters' absolute ages give useful information about the age of the Universe.

Carney tells the story like a thriller. Some topics are threaded through the story and recur on different occasions, possibly with a slightly different look. Carney's views on other topics are occasionally stated with perhaps too much confidence. But even the most cautious investigators will feel that at some point they know who the murderer is, and so must read to the last page to learn whether they are right or wrong.

Harris considers globular clusters from another perspective: how they can be used to trace the processes of parent galaxy formation and evolution in our own Galaxy and also, as far as they have been identified, in external galaxies. His prose and method are clear and rigorous, and he leads us from one logical step to the next and on to the conclusion.

### New Journals

This year, *Nature's* annual new journals review supplement will appear in the 18 October issue. Publishers and learned societies are invited to submit journals for review, as well as details of any eligible electronic journals, taking note of the following criteria:

- Journals must have first appeared during or after June 1999 and published at least four separate numbers by the end of June 2001.
- Journals covering any aspect of science are eligible, although those dealing with clinical medicine and pure mathematics are excluded, as are newsletters and publications of abstracts.
- Frequency of publication must be at least three times a year.
- The main language must be English.
- Deadline for submission is 5 July.

For each eligible title, please send at least four different issues (the first, the most recent and any two others), together with full details of subscription rates, to: Isobel Flanagan, *Nature*, Porters South, Crinan Street, London N1 9XW, UK. Tel: +44 (0)20 7843 4542. Fax: +44 (0)20 7843 4596/7. e-mail: i.flanagan@nature.com

He starts by analysing the globular cluster system in the Milky Way, and its chemical, dynamic and kinematic properties. From these we can deduce the mass of the Galaxy and speculate on how it formed. We read again how the distance scale for globular clusters has been defined and critically evaluated using many different methods, and about the globular cluster luminosity function, which is a powerful tool for estimating extragalactic distances. Finally, we go beyond our own Galaxy to investigate the properties of the globular cluster systems in external galaxies — from Andromeda (the nearest) to those in the Virgo and Fornax clusters of galaxies. As the data have improved, so too has our understanding of galaxy formation, even though this brings with it an increasing complexity in the observational evidence and interpretative models. It is a great pleasure to read about this fascinating journey.

Astrophysics evolves fast, and some of the topics described as open questions or given tentative answers are now more firmly or somewhat differently answered; for example, what the relative luminosities of the stars RR Lyrae and Cepheid have revealed about the distance scale. Other such issues are the ages of globular clusters, or the fraction of evolved blue stars as a function of age.

The unavoidable repetitions between the two sets of lectures, far from being a drawback, offer the advantage of independent views of the same topic and how varied they can be. Together, the lectures present a great deal of information on globular clusters which is supported by precise technical detail, all the basic relevant formulae, statistical information on the available data and a wealth of references. The book is sure to become a basic reference source for scholars in this field for years to come. ■

Carla Cacciari is at the Osservatorio Astronomico di Bologna, via Ranzani 1, 40127 Bologna, Italy.

● Another recent book on stars and their evolution is *Extreme Stars: At the Edge of Creation* by James Kaler (Cambridge University Press: 2001. 236 pp. £22.95, \$34.95), which looks at some of the extremes of stellar behaviour with the non-specialist reader in mind.

● *Cataclysmic Variable Stars: How and Why They Vary* by Coel Hellier (Springer-Praxis: 2001. 210 pp. £24, \$39.95) is an introductory account of these binary stars for those with a background in astronomy or astrophysics.

● *The Cambridge Dictionary of Space Technology* by Mark Williamson (Cambridge University Press: 2001. 464 pp. £27.95, \$39.95) has 2,300 entries from A-4 ("the original engineering designation for the V-2") to Zvezda ("a module of the International Space Station") and "lists fundamental terms that will remain in common usage for the foreseeable future".

## Science in culture

### The look of lightning

Contrary to conventional belief, lightning is never zig-zag.

Martin Kemp

What does lightning look like? The way to find out now seems obvious. Simply take an 'instantaneous' photograph. This was exactly the solution proposed in the early 1880s by the amateur photographer William Jennings, of Philadelphia. In a small notebook, now at the George Eastman House in Rochester, Jennings juxtaposed the varied, dendritic traces of actual lightning and thunderbolts with the stock portrayal by artists, which he called an "awkward zig-zag". He proudly proclaimed that "for over fifteen years he [Jennings] made lightning photographs in various parts of the world; no two of which were alike, and none 'zig-zag'".

The jagged shaft of conventional lightning still retains a strong popular hold, from Jupiter's hand-thrown thunderbolts, through warning signs of high voltage on pylons to logos of modern electricity companies. For showing the wrongness of the conventional zig-zag, Jennings was awarded the John Price Wetherill medal by the Franklin Institute — a particularly apposite award for a fellow Philadelphian of the great statesman and natural philosopher.

The truth was just as surely established by Eadweard Muybridge's famous serial photographs of galloping horses. The photographs, made for Governor Leland Stanford in California during the 1870s, had shown for ever the error of the artists' convention of a running horse with its four legs simultaneously reaching out in front and behind. The famous French painter of equestrian scenes, Ernest Meissonier, lamented that "all these years my eyes had deceived me. I find I have been wrong."

But are matters that simple? If we ask, "false for what?", the issue become less straightforward. The galloping horse's legs, like the flash of lightning, are too quick to be



'frozen' or seen in discernible sequence by the human eye. If a representation is to evoke the way we actually see a rapidly moving object, it is the 'instantaneous' photograph that becomes false. It is rather like saying that X-rays give the true picture of the human body rather than our customary vision of light on its exterior surface. If we can see any aspect of the galloping horse's legs, it is at the end of their travel, at the moment when all the legs are momentarily, if sequentially, stationary in front and behind.

Jennings's patterns of lightning are, if anything, even more problematic perceptually. Not only are the configurations infinitely varied, in a way that we now recognize as chaotic, but his photographic processes were actually rather slow compared to the lightning. What he captured was the course of the lightning over a definite time-span, through photography's equivalent of the persistence of vision. His process captured the linear trace that coursed from the heavens while the eye of his camera remained open and its retinal surface remained receptive.

We may readily recognize the forked menace of photographed lightning, just as the impressionist painters effectively exploited Muybridge's leg positions for a horse in rapid motion. However, the jagged zig-zag of convention still succeeds in working potently with the 'impression' produced when we stare at lightning using the perceptual apparatus with which we have been endowed. It also exudes a sense of graphic aggression that is highly expressive of lightning's destructive potential. We may know what lightning looks like, in an apparently objective sense, but how we actually see it is a very different matter. ■

Martin Kemp is in the Department of the History of Art, University of Oxford, 59 George Street, Oxford OX1 2BE, UK.  
The work of William Jennings can be seen in the exhibition *Light!*, which runs at the Carnegie Museum of Art in Pittsburgh until 29 July.



# Strings and things

Language is stretched to its limits in an effort to describe mathematics.

Jon Turney

“Common metaphors are the best because they are the only true ones,” said Argentinian poet Jorge Luis Borges. If, along with physicist Steven Weinberg, you dream of a final theory, maybe superstrings will be the answer. But how are you going to explain them? Well, says Weinberg in *Dreams of a Final Theory* (Pantheon, 1994), “these strings can be visualized as tiny one-dimensional rips in the smooth fabric of space”. But what does the fabric of space look like, and how does it get torn? Mixing metaphors like this (maybe the strings represent the fabric of space unravelling) betrays the strain new theory places on everyday language.

Science wants to be precise, unambiguous, logical and universal. Natural language is none of these things: hence scientists’ addiction to mathematics. But if they cannot manage with words alone, neither can they do without them. Words lend coherence, build understandable narratives, explain what the mathematics is about.

In physics, especially, the struggle to fathom what underlies appearances is always twofold. First, there is a search for a formalism that offers some kind of fit with reality. Then there’s a groping for a description in words which makes sense of the formalism.

And it is through words that the wider culture begins to make sense of what the scientists are trying to say. It may pick up on turns of phrase or key metaphors from professional discourse: big bang, black hole. But these are invariably enriched and elaborated in new ways as popular accounts of new theories proliferate.

Popular-science authors gradually build up a stock of explanations. While literary writers worth reading strive for freshness of expression, science writers are happy to re-use others’ tried and tested analogies. Deploying a genuinely new one requires a great deal of thought. Is it really fit for the purpose? So run-of-the-mill science books tend to use what are now the conventional non-mathematical explanations of quantum mechanics, say, or of Einsteinian space-time.

Right now, superstrings offer a fascinating demonstration of how hard it is to convey what the mathematics might be telling us before the conventional explanation has had time to bed down. Physics has coined the term as a shorthand for a new mathematical

entity. But how to go beyond the early, rather tentative, attempts, such as Weinberg’s, to elucidate?

In a way, superstring theory says there is only one thing to describe. The whole point of the theory is that the Universe contains just one fundamental kind of entity, which can manifest itself as all the particles and forces of conventional theory. But we want a non-mathematical sense of what kind of thing it is. Where to start? From its name we know that superstring theory will involve string-like entities. But what kind of string? A ball of garden twine, perhaps? A puppet string? A bow string? No: a violin string, vibrating under tension and generating harmonies.

But no sooner are we dusting off our intuitions about violin strings than the image undergoes violent alteration. For these strings can never be bowed or plucked. They are almost inconceivably small,  $10^{20}$  times smaller than an atomic nucleus. For Brian Greene — whose extended account of string theory in *The Elegant Universe* (Jonathan Cape, 2000) won last year’s British science book prize — strings appear as “tiny, one-dimensional filaments somewhat like infinitely thin rubber bands”. Nor are these strings stretched between fixed points like those in a musical instrument because, again like rubber bands, they are looped around.

Now forget the rubber. Each loop is not made of any ordinary material. They are “minute loops of energy”, says John Horgan in *The End of Science* (Addison-Wesley, 1996). “A string is just a loop drawn in space”, offers Lee Smolin, in *Three Roads to Quantum Gravity* (Weidenfeld & Nicolson, 2001). It is not made of material at all. It is what material is. As for elementary particles, there is no real content to the question, what are strings made of? They just are. Rather ask what they can do.

The main thing they can do is vibrate, oscillate or resonate — vibration is the most popular term, and harks back once more to violin strings. They generate tones by res-

If you cannot explain a result in physics in simple, non-technical terms, then you do not fully understand it.



Superstring: shorthand for a mathematical entity.

onating at frequencies that allow a whole number of waves to occupy the length of the string. Then, as Horgan puts it, particles with different masses and force charges arise from vibrations of elementary strings “just as vibrations of violin strings give rise to different notes”.

So far, the macro-metaphor is working well. The physicists’ strings may be vanishingly small, looped, and under fantastically high tension. But they still seem string-like.

Then, however, comes another property that doesn’t fit the word ‘string’ at all. For individual strings can merge or divide. A particle collision is now described as the merger of two strings to produce a third. This string travels a bit, and then releases the energy derived from the two initial strings by dissociating into two strings that travel onward. This is getting harder to picture, even when accompanied with diagrams that conform to Smolin’s suggestion that “when a string moves in time it makes a tube rather than a line”. But the notion of collision and merger remains completely unexpected from the macro-analogy which the term string most commonly invites.

At the moment, though, it is difficult to see where else a writer eschewing mathematics could turn for an intelligible way of approaching these strange objects. You can hardly begin with 11-dimensional membranes and supersymmetry, which is where the development of the theory eventually takes you. But as long as there are writers, such as Greene, who agree with Ernest Rutherford — that if you cannot explain some result in physics in simple, non-technical terms, then you do not fully understand its origin, meaning or implications — they will keep trying to stretch our intuitions about strings into new shapes. ■

Jon Turney is in the Department of Science and Technology Studies, University College London, Gower Street, London WC1E 6BT, UK.

# A tool, not a tyrant

Mott Greene

**I** believe many will discover in themselves a longing for mechanical explanation which has all the tenacity of original sin ... But nevertheless, just as the old monks struggled to subdue the flesh, so must the physicist struggle to subdue this sometimes nearly irresistible, but perfectly unjustifiable desire."

When Percy Bridgman wrote these words in the 1920s, he was preparing the way for a physics no longer unified by mechanics. Seventy-five years later, physicists still find it hard to abandon the search for mechanical explanation. Mechanisms and mechanics rule their minds and hearts: they still speak of quantum 'mechanics' even though there are no mechanics in it; the use of the word is a gesture of preferences and hope, without any particular physical meaning.

"Attempts to understand the motion of the electrons going around the nucleus by using mechanical laws," wrote Richard Feynman, "were a real failure: all kinds of predictions came out wrong."

The persistence of our desire for a mechanical explanation of every phenomenon may, however, not be only a preference or an ingrained habit of thought, but rather an injunction from nature, and the expression of our need to control, rather than merely to under-

stand, the world. The criteria we use to settle scientific questions are clearly constructed to favour solutions that have a chance of technological application. The search for a 'mechanism' may thus be part of a profound and ancient evolutionary strategy to help us survive, now built into our language and thought processes. There is more to our love of mechanism than our evolutionary past, however.

"I never satisfy myself until I can make a mechanical model of a thing," wrote Lord Kelvin, who required a mechanical picture as a metaphorical structure to guide his thinking. For him, a proper mechanical model was a tool of discovery. It was a structure of thought that would suggest new relations not at first apparent, and uncover aspects of reality not yet observed. Kelvin, and many before and after him, have spoken of the need for a mechanism when they meant to indicate that the business of explaining is incomplete until we discover the 'prime mover' in any physical situation. "Men are never satisfied," said Aristotle, "until they know the 'why' of a thing." To uncover a 'mechanism' is, for most of us, to move our ignorance of the world back one step towards that ever-elusive 'why'.

As the dictionary definition suggests, 'mechanism' is generally used as a synonym for 'cause', and the demand for a mechanism is an exhortation to pursue our inquiries into phenomena to the limits of our understanding.

## Mechanism

*"The machinery (lit. or fig.) by means of which some particular effect is produced." Oxford English Dictionary.*

For example, the hypothesis of continental drift was rejected in the 1920s for lack of such a mechanism, and its acceptance in the 1970s coincided with the development of the hypothesis of seafloor spreading, invoked as a mechanism for drift. Since then, inquiries into the subject have proceeded in search of the 'mechanism of the mechanism' — locating the cause of seafloor spreading in the convection of the Earth's mantle. Now, the search continues in an effort to locate a mechanism controlling the timing and location of convection: mechanisms within mechanisms, like nested Russian dolls.

The idea of mechanism is thus deeply lodged in us, and it serves us well, but we can place too much emphasis on it. This is an important point. Scientific theories (not mere hypotheses) based on abundant and convincing evidence and sound arguments have sometimes been forestalled for decades because they lacked a 'mechanism'. This was the fate of Darwin's theory for three-quarters of a century. The theory of evolution by natural selection was based on the existence of variation among members of a species, descent with modification from parent organisms, and differential survival of the offspring of different parents within a species — all of which could be easily observed. Darwin's inability to offer generally accepted mechanisms (causes) for these allowed not only his scientific opponents, but those theological opponents of any sort of evolution by any imaginable mechanism, to score debating points against the theory.

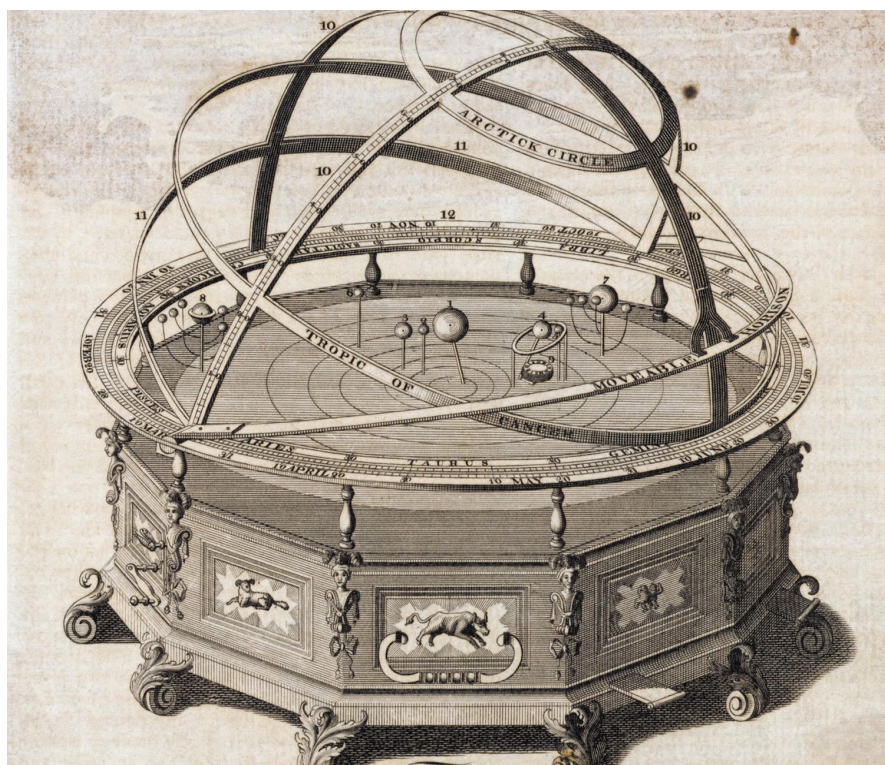
We might better say, with Bridgman, that for all its usefulness as a tool for thinking, our search for a mechanism must be regarded "no more seriously than is a mnemonic device, or any other artifice by which a man forces his mind to give him better service". We should also approach the question of mechanism with some humility, recalling Immanuel Kant's warning that "human reason ... is called upon to consider questions, which it cannot decline, as they are presented by its own nature, but which it cannot answer, because they transcend every faculty of mind."

Mott Greene is at the University of Puget Sound, Tacoma, Washington 98416, USA.

### FURTHER READING

Bridgman, P. W. *The Logic of Modern Physics* (Macmillan, New York, 1927).

Feynman, R. *The Character of Physical Law* (MIT Press, Cambridge, MA, 1965).



Mechanizing the planets: an orrery, a mechanical model of the Solar System.

BETTMANN/CORBIS



# The promise of protonics

Truls Norby

Fuel cells seem a good alternative to dirty and wasteful combustion. But known electrolytes — the crucial component — all have flaws. A new solid-state proton conductor suggests a solution.

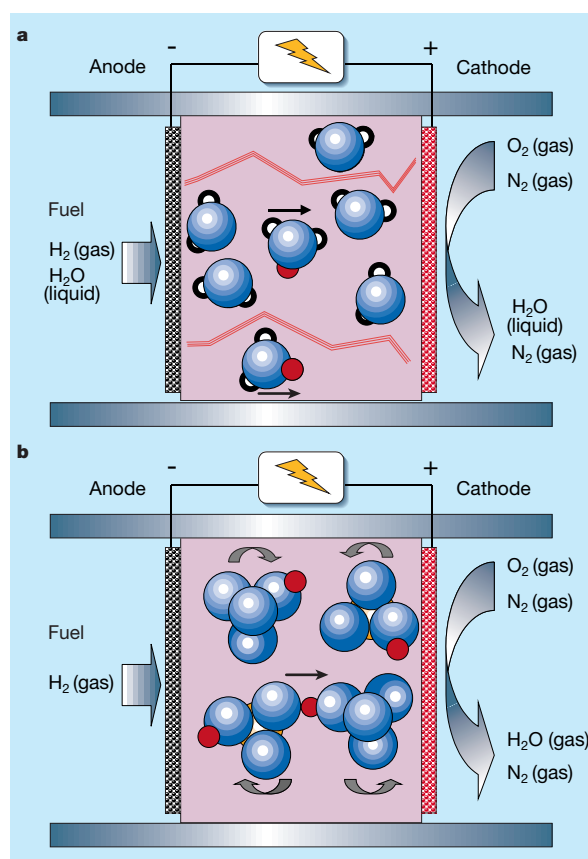
Fuel cells convert chemical energy directly to electrical energy cleanly and efficiently. But materials problems have delayed good commercial products. An essential component of a fuel cell is the electrolyte — a material that conducts ions. It can be aqueous, as in the phosphoric acid and alkaline fuel cells, or molten, as in the molten carbonate fuel cell; the mobile ions in these cells are  $\text{H}_3\text{O}^+$ ,  $\text{OH}^-$  and  $\text{CO}_3^{2-}$ , respectively.

A solid electrolyte is preferable because it allows sturdier, more efficient and corrosion-resistant systems to be built. The solid oxide fuel cell (SOFC) exploits the high mobility of oxygen ions,  $\text{O}^{2-}$ , in certain oxides at temperatures greater than 600 °C. Commercial SOFCs are candidates for stationary power installations and auxiliary power units in trucks. But problems with materials stability and performance at these high temperatures need to be solved. On page 910 of this issue, Haile *et al.*<sup>1</sup> describe a solid electrolyte based on a proton-conducting acidic salt, which operates in the much-sought-after intermediate temperature range.

At the lower end of the temperature scale, polymer electrolyte membranes that conduct protons ( $\text{H}^+$ ) go by many names and abbreviations: polymer electrolyte membranes (PEM), proton exchange membranes (also PEM), polymer electrolyte fuel cells (PEFCs) and so on. They are neither truly solid, nor do they only conduct protons: they contain liquid-like regions of water between the polymer chains (Fig. 1a). The protons can jump between the water molecules, but mainly ride on them as  $\text{H}_3\text{O}^+$  ions. Moreover, hydrogen bonding of additional water molecules to each proton creates drag, which slows down ion conduction. A bigger problem with these polymers is that they are fairly permeable to fuel molecules (such as hydrogen and methanol), thereby wasting energy by 'chemical short-circuit'.

To keep the water content high, PEFCs are operated below the boiling point of water, 100 °C. This requires the very best and most expensive of catalysts, platinum. And unlike SOFCs, they give off 'waste' heat at too low a temperature to be useful in, for example, domestic heating. But the flexibility and sturdiness of polymer-based fuel cells make them attractive for mobile applications. The PEFC will no doubt be used in the first fleets

Figure 1 Fuel cells that use proton conduction. a, Polymer electrolyte fuel cell. The polymer membrane consists of polymer threads interspersed with aqueous regions, where  $\text{H}_3\text{O}^+$  ion conduction — a proton (red) attached to a water molecule — occurs. Water molecules are a problem because they slow ion transfer, and require high water content in both the fuel and exhaust electrodes, which limits the operating temperature to below 100 °C. b, Fuel cell with solid-acid ( $\text{CsHSO}_4$ ) electrolyte. At intermediate operating temperatures (150 °C to 200 °C) the loose lattice of rotating oxyanions,  $\text{SO}_4^{2-}$  ions, allows proton transfer as indicated. Haile *et al.*<sup>1</sup> have created a laboratory fuel cell using  $\text{CsHSO}_4$  as the electrolyte, which operates stably between 150 °C and 160 °C.



of cars powered by fuel cells. New water-replacement systems, such as imidazole ( $\text{C}_3\text{H}_4\text{N}_2$ ), are currently being investigated in order to increase the operating temperatures of polymer electrolytes.

Given the problems with existing fuel cells, an ionic conductor working at intermediate temperatures (100–600 °C) is an ultimate goal of solid electrolyte research (often called ionics). Among the best candidates are the true proton conductors. In these there is usually no water, only lone protons bonded to oxygen ions as  $\text{OH}^-$  groups. Such compounds include hydroxides (for example,  $\text{NaOH}$ ,  $\text{Mg}(\text{OH})_2$  and  $\text{FeOOH}$ ) and acidic salts (such as  $\text{KHSO}_4$  and  $\text{KH}_2\text{PO}_4$ ). Unfortunately, most are poor proton conductors because the protons are too strongly bonded to their host oxygen ions, and they decompose or melt before the protons become mobilized.

A quarter of a century ago, a few of the acidic salts, called solid acids, were known to be good proton conductors (see ref. 2 for

a review). Solid acids are compounds, such as  $\text{KHSO}_4$  and  $\text{CsHSO}_4$ , whose chemistry and properties lie between those of a normal acid (such as  $\text{H}_2\text{SO}_4$ ) and a normal salt (such as  $\text{K}_2\text{SO}_4$ ). They usually consist of oxyanions, such as  $\text{SO}_4^{2-}$ , that are linked together by hydrogen bonds. It was not until the 1980s that  $\text{CsHSO}_4$  was found to undergo a rather spectacular phase transition at about 140 °C during which the proton conductivity increases by almost four orders of magnitude<sup>3</sup>.

In this higher-temperature, so-called superionic phase, the structure of  $\text{CsHSO}_4$  loosens up to let the  $\text{SO}_4^{2-}$  ions rotate so that a proton on one of the oxygen ions can approach an oxygen ion on a neighbouring oxyanion, facilitating proton transfer (Fig. 1b). A structural chemist would say that the four available locations for a proton on each oxyanion have become equivalent, so that the proton can choose to occupy any one of the four. And with these protons sitting on the liberated  $\text{SO}_4^{2-}$  groups around easily

polarizable  $\text{Cs}^+$  ions, you have the recipe for a good ionic conductor.

For a couple of decades,  $\text{CsHSO}_4$  has been the model high-temperature proton conductor. But no one tested it in a real fuel cell. They may have been discouraged by its softness in the superionic phase, its solubility in water, the difficulty of synthesizing the compound with electrodes attached and the narrow operating window available (between 140 °C and its melting/decomposition temperature above 200 °C). Haile *et al.*<sup>1</sup> have now responded to the increasing clamour for intermediate-temperature solid electrolytes, and proton conductors in particular, and tried out  $\text{CsHSO}_4$  in a laboratory set-up.

They made a fuel cell by sandwiching the solid acid between two electrode layers made from  $\text{CsHSO}_4$  powder mixed with platinum and carbon and a volatile organic. The entire fuel cell was compressed to provide good contact between the electrolyte and electrode layers. On heating, the organic component evaporated, leaving behind a porous electrode structure with the possibility of multiple contacts between the fuel (or oxygen, at the other end of the cell), the electrolyte and the electrodes.

Next came the question of whether the solid acid electrolyte was stable when exposed to the operating conditions of the fuel cell: a temperature of 160 °C, hydrogen gas at one electrode and oxygen at the other. The laboratory fuel cell performed stably over several days, burning the world's cleanest fuel — hydrogen — and emitting the most environmentally friendly effluent, pure water.

Haile and her colleagues<sup>1</sup> have been studying for some time another class of proton conductors, in which the protons are not part of the structure, but merely defects dissolved in water-containing atmospheres. The breakthrough for these materials came in the early 1980s when Iwahara and co-workers showed that chemically doped  $\text{SrCeO}_3$  and other perovskite-related oxides became proton-conducting in humid atmospheres and at high temperatures up to 800 °C. Iwahara demonstrated fuel cells and other devices based on a number of perovskite-related oxides, and conceived the term protonics for the rapidly expanding field of solid-state proton conductors<sup>4</sup>.

These high-temperature perovskite conductors lose protons at high temperatures, so the optimum temperature of proton conduction is a compromise between proton concentration and mobility. The proton conductivity typically peaks at around 400–600 °C. From these temperatures and down to those of solid acids such as  $\text{CsHSO}_4$  (160 °C) there is a gap in fuel-cell temperatures for which, as yet, there is no real candidate for a solid electrolyte. This is an important temperature range because it should allow good materials stability, fast

reaction kinetics, and manageable heat recovery for a wide variety of applications.

Haile *et al.*<sup>1</sup> are applying for a patent for fuel cells based on solid acid salts, but are otherwise realistically modest about their results. The  $\text{CsHSO}_4$  electrolyte they used was a millimetre or more thick, whereas real applications will require micrometre-thin films to reduce the resistance of the electrolyte layer. The performance of the electrodes may also be of concern; slow charge transfer appears to cause further drops in voltage. Overall, the laboratory fuel cell delivers a modest 44 milliwatts per square centimetre, which compares unfavourably with the output of state-of-the-art PEFCs and SOFCs. More importantly, one can imagine that operation times of tens of thousands of hours will reveal degradation of the electrolyte due to creep, evaporation and reduction of sulphate in the solid acid by the fuel. From a practical

point of view, accidental overheating would melt or decompose the electrolyte and water flooding would easily dissolve it.

Nonetheless, Haile *et al.* have proved the concept; true solid-state proton conductors can indeed be put to work in fuel cells, and the relevance of protonics has been extended into the intermediate temperature range. If  $\text{CsHSO}_4$  is not the most practical electrolyte, then there is a rich chemistry of solid acids and hydroxides with much yet to discover. ■

Truls Norby is in the Department of Chemistry, University of Oslo, Centre for Materials Science, Gaustadalleen 21, NO-0349 Oslo, Norway.

e-mail: truls.norby@kjemi.uio.no

1. Haile, S. M., Boysen, D. A., Chisholm, C. R. I. & Merle, R. B. *Nature* **410**, 910–913 (2001).
2. Glasser, L. *Chem. Rev.* **75**, 21–65 (1975).
3. Baranov, A. I. *et al.* *JETP Lett.* **36**, 459–462 (1982).
4. Iwahara, H. in *Proc. 17th Riso Int. Symp. Mat. Sci.* (eds Poulsen, F. W. *et al.*) 13–28 (Riso National Laboratories, Roskilde, Denmark, 1996).

#### Molecular motors

## Doing a rotary two-step

Mark J. Schnitzer

By spinning around, the ATP synthase converts energy from electrochemical to chemical form for storage. A cunning assay reveals intricate details of the rotation.

How do organisms power themselves? This simple question marked the start of a decades-long quest that led eventually to a rotary enzyme called ATP synthase. Using energy derived from proton gradients created during photosynthesis or respiration, this enzyme produces ATP (adenosine triphosphate) molecules — the main cellular energy store. A challenge now is to understand how ATP synthase works. It is thought to contain two rotary motors, called  $F_0$  and  $F_1$ , which are rotationally coupled through their drive shafts. This device uses motion, in the form of 120° rotary steps by  $F_1$ , as an intermediate, converting energy from electrochemical through mechanical to chemical form<sup>1</sup>. On page 898 of this issue<sup>2</sup>, Yasuda and colleagues describe their elegant studies of the  $F_1$  motor, and show that each rotary step actually consists of two smaller substeps. More generally, their work highlights some central issues concerning energy flow in proteins.

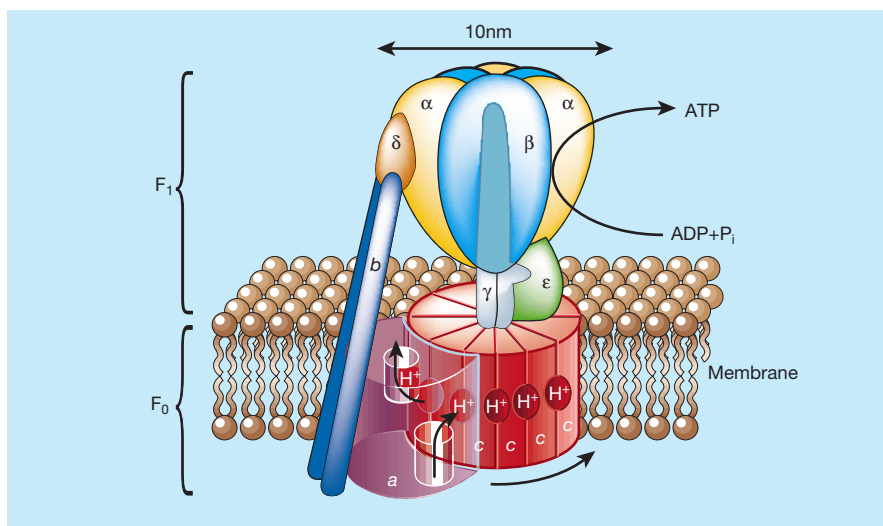
The  $F_0$  and  $F_1$  motors have specialized functions (Fig. 1). The  $F_0$  motor is bound to membranes of energy-generating cellular structures such as mitochondria, and channels protons through its rotor and non-rotating stator to drive rotation. Meanwhile, the stator of  $F_1$  catalyses the production of ATP from ADP (adenosine diphosphate) and inorganic phosphate, provided that  $F_0$  drives the  $F_1$  rotor with sufficient torque

(rotary force). The catalytic sites of  $F_1$  are within the three so-called  $\beta$ -subunits, which are arranged with tripartite symmetry in the stator. Both  $F_0$  and  $F_1$  can operate in reverse, and isolated  $F_1$  is called the  $F_1$ -ATPase because, in the absence of torque from  $F_0$ , it hydrolyses ATP to power reverse rotation.

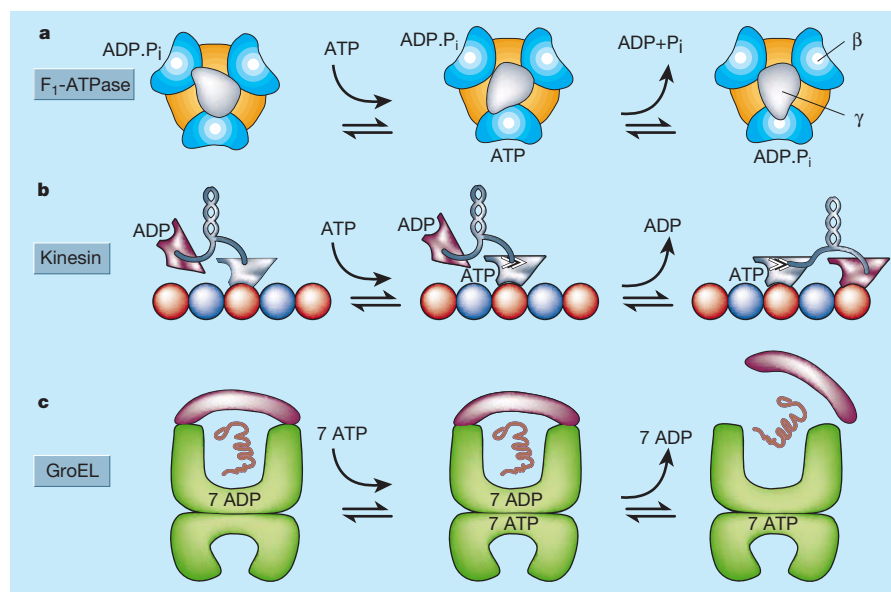
There is only indirect evidence that  $F_0$  can drive rotation<sup>3–5</sup>. But rotation of  $F_1$  sub-complexes has been visualized directly *in vitro*, by fixing the stator to a glass surface and using a micrometre-sized filament attached to the rotor's  $\gamma$ -subunit as a marker<sup>6,7</sup>. With this assay, 120° rotary steps can be detected by monitoring the movement of the filament at low ATP concentrations, such that the binding of ATP to  $F_1$  is the rate-limiting event between steps<sup>7</sup>. The distribution of times between steps is consistent with one ATP molecule being hydrolysed per step.

Once a step begins, the time taken for the 120° turn depends on the frictional drag imposed by the filament and so on the filament's length. Drag also smoothes out any rapid fine structure within the protein's 120° movement when monitored via the filament. This obscures how rotor movements are linked to hydrolysis events, which probably occur in reverse during ATP synthesis. So, to break the speed barrier imposed by the filament, Yasuda *et al.*<sup>2</sup> have revved up the assay by using a bead with a diameter of a mere 40 nanometres as the marker. Using





**Figure 1** ATP synthase — energy converter. The enzyme consists of two motors,  $F_0$  and  $F_1$ . The transmembrane  $F_0$  motor has one  $a$ , two  $b$  and nine to twelve  $c$  subunits. The soluble  $F_1$  motor has three  $\alpha$  and three  $\beta$  subunits, and one each of the  $\epsilon$ ,  $\delta$  and  $\gamma$  subunits. The  $F_0$  rotor (consisting of the  $c$  subunits) and the  $F_1$  rotor ( $\epsilon$  and  $\gamma$  subunits) are connected, as are the respective stators (the  $F_0$  stator comprises the  $a$  and  $b$  subunits; the  $F_1$  stator consists of the  $\delta$  subunit and the three  $\alpha$  and three  $\beta$  subunits). During ATP synthesis,  $F_0$  channels protons across the membrane to drive rotation, which in  $F_1$  occurs as  $120^\circ$  rotational steps. Yasuda *et al.*<sup>2</sup> show that each  $120^\circ$  step is composed of two smaller substeps of  $90^\circ$  and  $30^\circ$ . Modified from refs 14, 15.



**Figure 2** Molecular machines that use ATP binding at one catalytic site to trigger a large conformational change and the release of ADP from another catalytic site.  $P_i$  is inorganic phosphate. **a**,  $F_1$ -ATPase, with the catalytic  $\beta$ -subunits in dark blue, the  $\alpha$ -subunits in yellow, and the  $\gamma$ -subunit in light blue. When ATP binds to one  $\beta$ -subunit, the  $\gamma$ -subunit rotates by  $90^\circ$ . ADP and  $P_i$  are released from another  $\beta$ -subunit, and the  $\gamma$ -subunit rotates a further  $30^\circ$ . **b**, Kinesin, a linear motor with two catalytic heads (purple and grey) that move along a microtubule filament (red and blue). ATP binding to the front head triggers a conformational change (shown by double arrowheads) in its adjoining neck linker domain<sup>9</sup> and induces the rear head to step forward and to release ADP. **c**, The chaperone GroEL (green), its co-chaperone GroES (purple) and a substrate protein (squiggly line). GroES acts as a cap, containing the substrate protein within GroEL while the substrate folds. GroEL has two rings, each of which bind seven ATP or ADP molecules. Cooperative binding of seven ATP to one ring triggers the other ring to release GroES, the substrate protein and seven ADP molecules.

a high-speed camera to capture footage of bead rotation, the authors find that, when attached to the  $\gamma$ -subunit off the rotational axis, the bead spins in an eccentric orbit during ATP hydrolysis. Rotational frictional

drag is so slight that, at saturating ATP concentrations,  $120^\circ$  steps begin roughly every 2 milliseconds, and an entire step is complete within about 0.25 milliseconds.

A potential problem with these rapid

speeds is that they might make it hard to detect substeps within a step, given that the camera acquires data only twice during a  $120^\circ$  step. However, it was known through the Nobel-Prize-winning work of Paul Boyer<sup>8</sup> that a large fraction of the energy liberated during the ATP-hydrolysis cycle is released when ATP binds tightly to one catalytic site. Most of the remaining energy is freed about 2 milliseconds later, when the products of hydrolysis (ADP and inorganic phosphate) are released from another catalytic site (Fig. 2a). It seemed plausible that these large energy changes might be accompanied by large conformational changes in  $F_1$ . If those changes involved rotational substeps, then it might prove possible to increase the average time between successive substeps to at least 2 milliseconds. The trick would be to pause rotation after the product-release substep, by lowering the ATP concentration enough to delay the next ATP-binding substep by 2 milliseconds or more. So, each substep would now follow a pause.

Yasuda *et al.*<sup>2</sup> adopt this tactic, and records of motion at  $20 \mu\text{M}$  ATP or below reveal  $90 \pm 10^\circ$  and  $30 \pm 10^\circ$  substeps, separated by about 2 milliseconds. The  $90^\circ$  movement must be the binding substep — its mean initiation time following a  $30^\circ$  movement depends on ATP concentration. Histograms of initiation times reveal the delay between  $90^\circ$  and  $30^\circ$  rotations, and show a time distribution consistent with two kinetic transitions, each requiring about 1 millisecond. A remaining challenge is to correlate these transitions and the substeps with structural forms of the catalytic  $\beta$ -subunits.

ATP synthase was one of the first enzymes known to show cooperative interactions between its catalytic sites. We now have explicit proof<sup>2</sup> that ATP binding to  $F_1$  triggers rotation.  $F_1$  thereby joins a growing list of molecular machines that use ATP binding at one catalytic site to induce product release from another catalytic site through large mechanical movements (Fig. 2b, c). Two other examples are the motor protein kinesin and the chaperone molecule GroEL, which helps newly formed protein chains to fold into the correct shape. Kinesin has two catalytic head regions that somehow allow the molecule to advance in discrete 8-nanometre steps. ATP binding at one head triggers a conformational change in the protein, allowing ADP to be released from the other head<sup>9</sup> and stepping to occur. GroEL has two ring domains; binding of seven ATP molecules to one ring triggers the other ring to release the substrate protein<sup>10</sup> and seven ADP molecules.

In all of these proteins, how does energy flow outwards from an ATP-binding site to initiate movement and product release? Energy must travel between catalytic sites — about 5 nanometres in  $F_1$ -ATPase<sup>11</sup>, some 9 nanometres in GroEL<sup>10</sup>, and potentially

up to 16 nanometres in kinesin<sup>9</sup>. Detailed modelling<sup>12</sup> of  $F_1$  indicates that ATP binding might trigger elastic bending in the catalytic  $\beta$ -subunit, which then applies torque to the  $\gamma$ -subunit in the rotor. In kinesin, a rigid helical structure may act as a mechanical relay between a phosphate-sensing loop at the ATP/ADP-binding site and a linker region that is usually mobile but which becomes immobilized upon ATP binding<sup>9</sup>. Energy flow in GroEL is poorly understood. Indeed, questions about all three enzymes remain; for example, do mechanical signals generally occur as elastic strain, as in  $F_1$ , or can some protein domains become thermally activated upon ATP binding? Temperature-dependent kinetic studies and efficiency considerations may help to separate entropic and enthalpic contributions, but energy flow will probably remain murky until investigated further. The hope is that new experimental approaches, such as those relying

on protein databases to enable statistical comparisons of closely related proteins<sup>13</sup>, will speed our understanding.

Mark J. Schnitzer is in the Biological Computation Research Department, Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey 07974, USA. e-mail: schnitzer@lucent.com

1. Boyer, P. D. *Annu. Rev. Biochem.* **66**, 717–749 (1997).
2. Yasuda, R., Noji, H., Yoshida, M., Kinosita, K. Jr & Itoh, H. *Nature* **410**, 898–904 (2001).
3. Dimroth, P. *Biochim. Biophys. Acta* **1458**, 374–386 (2000).
4. Sambongi, Y. *et al. Science* **286**, 1722–1724 (1999).
5. Rastogi, V. K. & Girvin, M. E. *Nature* **402**, 263–268 (1999).
6. Noji, H., Yasuda, R., Yoshida, M. & Kinosita, K. *Nature* **386**, 299–302 (1997).
7. Yasuda, R., Noji, H., Kinosita, K. & Yoshida, M. *Cell* **93**, 1117–1124 (1998).
8. Boyer, P. D. *Biochim. Biophys. Acta* **1140**, 215–250 (1993).
9. Vale, R. D. & Milligan, R. A. *Science* **288**, 88–95 (2000).
10. Sigler, P. B. *et al. Annu. Rev. Biochem.* **67**, 581–608 (1998).
11. Abrahams, J. P., Leslie, A. G., Lutter, R. & Walker, J. E. *Nature* **370**, 621–628 (1994).
12. Wang, H. & Oster, G. *Nature* **396**, 279–282 (1998).
13. Lockless, S. W. & Ranganathan, R. *Science* **286**, 295–299 (1999).
14. Elston, T., Wang, H. & Oster, G. *Nature* **391**, 510–513 (1998).
15. Boyer, P. D. *Nature* **402**, 247–249 (1999).

## Fundamental physics

# Newton rules (for now)

Frank Wilczek

Three hundred years after Newton explained the falling of an apple and the motion of the planets, physicists are beginning to test his universal law of gravity down to micrometre distances — with interesting results.

Newton's formula for the force of gravitational attraction is the oldest, and perhaps the most revered, universal law of physics. Yet it is not sacred. The experimental support for Newton's force law at sub-centimetre distances is surprisingly weak. There are theoretical suggestions about why and how deviations from Newton's law might arise at these distances. A group of experimentalists at the University of Washington, Seattle, has tackled the question, using ingenious but surprisingly low-tech, small-science techniques to test newtonian gravity down to 200  $\mu\text{m}$ . As they report in *Physical Review Letters*<sup>1</sup>, Newton's law still rules, which puts some theoretical speculation about large 'extra dimensions' on the ropes.

In modern physics, the precise form of Newton's formula for the gravitational force — proportional to mass, inversely proportional to distance squared — is rooted in profound principles. Force proportional to mass means acceleration independent of mass. Thus Newton's law implies that the motion induced by gravity is independent of the material properties of the body it acts upon. This idea is deeply engrained in Einstein's theory of general relativity, in which gravitational fields (produced by matter) change the geometry of space-time, causing it to become curved. It is the curvature of space-time that controls the natural accel-

eration of bodies. The inverse square form of Newton's law is a consequence of the field equations for gravity. The form of these equations, in turn, is dictated by the general principles of quantum field theory (specifically, the requirement that they derive from a local lagrangian function).

How, then, might deviations arise? One way is through the existence of extra, non-gravitational forces. Indeed, we know that there are other forces in physics, even at macroscopic distances. For example, there are  $1/r^2$  electric forces between charged bodies, and  $1/r^4$  magnetic forces (with complicated angular dependence) between magnetic bodies, where  $r$  is the separation. Even between unscreened neutral, non-magnetic bodies there are attractive short-range  $1/r^7$  van der Waals forces. If our goal is to test the foundations of physics, however, our focus must be on non-electric, non-magnetic bodies, and forces beyond van der Waals.

Within the framework of quantum field theory, forces are associated with the exchange of virtual particles. Thus new forces are associated with new particles. The range of the force depends inversely on the mass of the particle. If the particle mass is  $m$ , then the range of the corresponding force is  $\hbar/mc$ , where  $\hbar$  is Planck's constant and  $c$  is the speed of light. Numerically, a particle of mass  $2 \times 10^{-5}$  electron volts (25 billion times lighter than the electron) yields a force

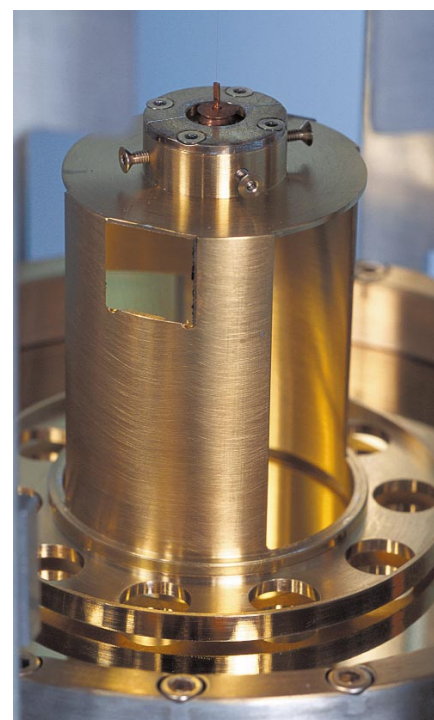


Figure 1 Apparatus used at the University of Washington<sup>1</sup> to test Newton's inverse square law. They measured the motion of a metal ring suspended above a rotating disk to show that Newton is correct down to ring-disk separations of at least 200  $\mu\text{m}$ . The photograph shows the metal ring, which has ten holes bored in it, suspended from a torsion pendulum. The rotating disk below the metal ring is hidden from view by a sheet of metal foil that minimizes electrical forces on the pendulum. As the disk rotates, the gravitational force causes the ring to twist back and forth ten times for each revolution of the disk. The twisting of the pendulum is measured by reflecting a laser beam from a small mirror attached to the upper part of the pendulum. The entire apparatus is coated in gold to prevent electrical forces interfering with the weak gravitational signal.

with a range of 1 centimetre. Thus in looking for deviations from Newton's force law — specifically new forces operating at super-molecular but sub-centimetre distances — we are probing for the existence of extremely light, extremely weakly interacting particles<sup>2,3</sup>.

Many examples of such particles have been suggested, with various motivations. So-called axions<sup>4</sup> arise in attempts to explain the accurate time-reversal symmetry of the 'strong force' — the short-range interaction that holds protons and neutrons together in nuclei. Dilatons arise in superstring theory (a fundamental theory of elementary particles that requires extra dimensions). Familons and modulons arise in attempts to understand the origin of the subtle differences between particles in different families — for example the difference between the electron and its heavier cousins, the muon and tau leptons<sup>5</sup>.

None of these exotic particles has yet been



detected. Physicists have seen only the effects of exchanging massless 'gravitons', which are the particles thought to be responsible for transmitting gravity — a force they already know exists. In 1998, Arkani-Hamed and colleagues predicted<sup>6</sup> new effects involving extra dimensions and gravity, which captured the imagination of many physicists, and heightened interest in the search for deviations from Newton's force law.

Arkani-Hamed *et al.*<sup>6</sup> proposed that gravitons sense not only the four (three of space and one of time) macroscopic dimensions available to other particles, but also one or more additional spatial dimensions. Although these extra microscopic dimensions were once assumed to be too small to see, in some models they might be as big as a millimetre. If the size of the extra dimensions is  $d$ , strong deviations from Newton's force law are predicted for separations less than  $d$ . In the language of particle physics, when gravity spreads to the extra dimensions it produces a whole tower of new particles — so-called Kaluza–Klein gravitons — that modify the inverse square law.

In their experiment at the University of Washington, Hoyle *et al.*<sup>1</sup> suspended a metal ring containing ten equally spaced holes from a torsion (twisting) pendulum (Fig. 1). The pendulum was placed over a cleverly shaped rotating disk that contained two layers of holes (ten in each layer). They measured the torque on the pendulum produced by rotating the disk at different frequencies. The torque was created by the tendency of gravity to align the holes in the ring and disk. The measurements were repeated at different ring–disk separations, going from 10 mm down to 200  $\mu\text{m}$ . All the components had to be precisely machined and electrically inert.

Hoyle *et al.*'s results show no deviations from newtonian physics, although the experiments were sensitive to forces of finite range and strength comparable to gravity down to 2 mm, and had significant sensitivity to stronger forces that kick in below this distance. The coupling strength of new force-transmitting particles with masses less than  $10^{-3}$  electron volts is therefore severely constrained by this result. It indicates that if the extra dimensions proposed by Arkani-Hamed *et al.*<sup>6</sup> exist, then they must be smaller than 200  $\mu\text{m}$ .

In the near future, the University of Washington group promises still more precise results, extending to even shorter distances. Another direction for this sort of experiment is to test whether the gravitational force is really independent of the type of material. Whereas gravity is strictly dependent on mass, other hypothetical particles will generally have some dependence on a combination of mass and other quantum numbers, such as spin. For example, axion-mediated forces depend primarily on

spin<sup>2</sup>. It is important to continue progress at this unusual frontier of fundamental physics experiments, which complements work using gigantic particle accelerators, is relatively cheap and is largely unexplored. ■

Frank Wilczek is at the Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139-4307, USA.

e-mail: wilczek@mit.edu

1. Hoyle, C. D. *et al.* *Phys. Rev. Lett.* **86**, 1418–1421 (2001).
2. Moody, J. & Wilczek, F. *Phys. Rev. D* **30**, 130–138 (1984).
3. Adelberger, E. *et al.* *Annu. Rev. Nucl. Part. Phys.* **41**, 269–320 (1991).
4. Kim, J. *Phys. Rep.* **150**, 1–177 (1987).
5. Long, G., Chang, H. & Price, J. *Nucl. Phys. B* **529**, 23–34 (1999).
6. Arkani-Hamed, N., Dimopoulos, S. & Dvali, G. *Phys. Lett. B* **436**, 257–263 (1998).

## Cell biology

# Proteolytic relay comes to an end

Martin Scheffner and Noel J. Whitaker

The cellular protein-degrading machinery detects target proteins by the signals they display. One such signal works for artificial test proteins, but only now has a natural equivalent been found.

The selective degradation of proteins is vital in every cell and organism. It is, for instance, involved in eliminating misfolded proteins and in controlling protein activity. Not surprisingly, then, malfunctions in protein degradation are associated with many human diseases, including neurodegenerative disorders and cancer. Proteins that are doomed to be destroyed exhibit distinct signals that mark them as targets for protein-degrading (proteolytic) systems. One such signal is the nature of the amino acid at the amino-terminal end ('N-end') of a target protein. This 'N-end rule' applies both to bacteria and to higher organisms, yet its physiological relevance has remained unknown, until now.

On page 955 of this issue, Rao *et al.*<sup>1</sup> report that site-specific cleavage of the SCC1 protein — an event that is necessary to allow duplicated chromosomes to segregate during cell division — generates a protein fragment that is targeted for destruction through recognition of its amino terminus. Moreover, failure to destroy this fragment results in chromosomal instability. So, one physiological function of the N-end rule is to rid the cell of the potentially harmful leftovers of site-specific protein cleavage.

The basis of selective protein turnover is the specificity of the interaction between a given proteolytic system and its target proteins. This implies that the targets must feature amino-acid sequences or structures that are specifically recognized as degradation signals by components of the proteolytic system. In a quest to identify such signals, Bachmair *et al.*<sup>2</sup> discovered that the type of amino acid present at the amino terminus of an artificial protein has a marked effect on the protein's turnover rate in budding yeast, *Saccharomyces cerevisiae*. According to the N-end rule, the amino-terminal amino acid can be characterized as having either a stabilizing or a destabilizing effect on a test protein. 'N-end-rule substrates' are those with amino-terminal amino acids that are destabilizing.

Members of the same group<sup>3</sup> went on to show that, in higher organisms, N-end-rule substrates are degraded by the so-called ubiquitin/proteasome pathway<sup>4</sup>. In this proteolytic system, substrates are first tagged with several copies of the small protein ubiquitin. Tagged substrates are then degraded by the 26S proteasome complex. The implication is that substrate recognition in this system is mediated mainly by a component of the ubiquitin-conjugating machinery — the protein UBR1, in the case of N-end-rule substrates<sup>5,6</sup>. Cells from bacteria to humans can process artificial proteins according to the N-end rule, with some variations in the amino-terminal amino acids that are stabilizing or destabilizing (see ref. 7). The actual mode of recognition, however, is probably different, because bacteria do not contain ubiquitin.

This is a fascinating story, but one with a significant caveat. Until now, there were no known cellular substrates — and thus no physiological functions — for the N-end rule. In fact, all known cytosolic and nuclear proteins in yeast have a stabilizing amino acid at their amino terminus. Does this mean that cells have evolved mechanisms to counteract the N-end rule? Or is there nonetheless a definable physiological role for this rule? A first clue to the answers came from studies of pathogenic viruses and microbes. These organisms frequently produce long polypeptide chains that are cleaved by site-specific proteases to generate distinct, active proteins. In some cases, this results in destabilizing amino acids being uncovered at the amino termini of the newly created proteins<sup>8–10</sup>. So, the products of site-specific proteases might be a source of protein targets for the N-end rule. This supposition has now been elegantly proved by Rao and colleagues<sup>1</sup>.

During the cell-division cycle, a cell must duplicate its chromosomes; the two copies of each chromosome are called sister chromatids. These must be held together until a later phase in the cell cycle (anaphase), at which

point they are segregated<sup>11</sup>. A complex called cohesin, one component of which is the SCC1 protein, holds chromatids together. As the cell proceeds into anaphase, the site-specific protease ESP1 is activated (Fig. 1) and cleaves SCC1 at two sites, allowing separation of sister chromatids<sup>12,13</sup>. One of the protein fragments generated by this cleavage has an arginine residue — a strongly destabilizing amino acid — at its amino terminus. In a careful biochemical and genetic analysis, Rao and colleagues provide compelling evidence that, in *S. cerevisiae*, this fragment (Arg-SCC1) is degraded in a UBR1- and N-end-dependent manner. Tellingly, this fragment accumulates in yeast cells that lack the *UBR1* gene, and replacement of arginine with stabilizing residues results in a long-lived protein.

Any perturbation of chromosome segregation could result in daughter cells that have an irregular number of chromosomes (aneuploidy) — a hallmark of human cancer. Rao and colleagues have found that *UBR1*-deficient yeast have an increased frequency of chromosome loss, which results from stabilization of Arg-SCC1. So, N-end-mediated degradation of Arg-SCC1 is important in ensuring chromosome stability in yeast, although the mechanism by which accumulation of Arg-SCC1 interferes with proper chromosome segregation remains unknown. Is this also true for humans? If so, are mutations in components of the N-end-rule pathway associated with cancer? We first need to know whether cleavage of the human counterpart of SCC1 also generates a protein with a destabilizing amino acid at its amino terminus (the sites of ESP1-mediated cleavage in

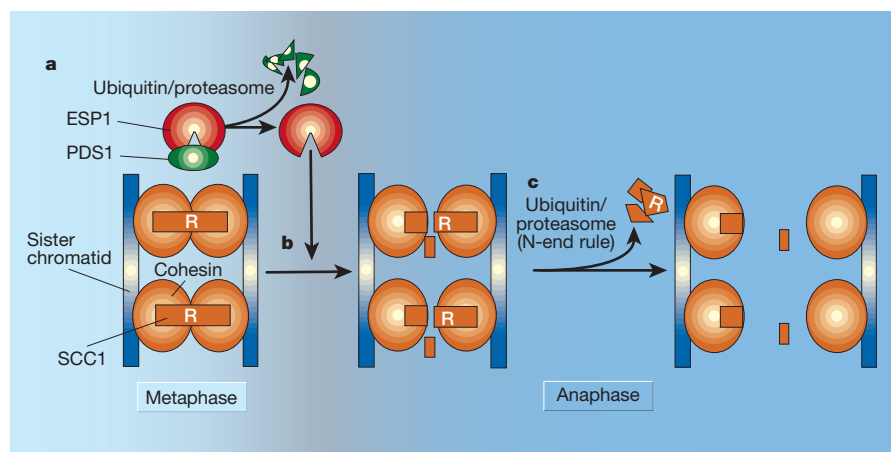
yeast SCC1 do not seem to be present in the human counterpart). It will also be interesting to study genetically engineered *UBR1*-deficient mice. If the N-end rule is indeed important for chromosome segregation in mammals, then such mice might be more likely to develop cancer.

The control of chromosome segregation is a beautiful example of the requirement for selective protein degradation in a temporally regulated process (Fig. 1). First, the inhibitor of ESP1 (called PDS1) is inactivated by ubiquitin/proteasome-mediated degradation<sup>14</sup>. This allows ESP1 to cleave SCC1, thereby freeing sister chromatids from each other<sup>12,13</sup>. Finally, the newly generated Arg-SCC1 protein is destroyed by the ubiquitin/proteasome system according to the N-end rule.

Several questions concerning this relay of proteolytic events remain. What happens to the other fragments of SCC1 that are produced by ESP1-mediated cleavage? Do these fragments have a function, and how are they finally destroyed? More generally, are any other proteins that are produced by site-specific proteases completely degraded as a result of recognition of their amino termini?

Whatever the answers, these findings also remind us that the total number of genes in a species' genome allows only a rough estimate of its total number of proteins. Nature has evolved several strategies to increase the protein repertoire, from differential splicing or editing of messenger RNA molecules to proteolytic processing of proteins.

Martin Scheffner is at the Institute of Biochemistry I, Faculty of Medicine, University of Cologne, Joseph-



**Figure 1** The role of protein breakdown in separation of sister chromatids during cell division in yeast. Sister-chromatid separation begins at the transition between the metaphase and anaphase of the cell-division cycle. Two protein-degradation (proteolytic) events are required. **a**, At the onset of anaphase, PDS1 — an inhibitor of the site-specific protease ESP1 — is destroyed by the ubiquitin/proteasome system. **b**, Liberated ESP1 then cleaves SCC1, a component of the cohesin complex that holds sister chromatids together. **c**, Rao *et al.*<sup>1</sup> show that the newly formed carboxy-terminal fragment of SCC1, which has an arginine residue (R) at its extreme amino terminus, is degraded by the ubiquitin/proteasome system according to the 'N-end rule'. The amino terminus of the protein fragment thus acts as a recognition site for the ubiquitin/proteasome system. This degradation is physiologically meaningful — if it is interfered with, division of the resulting daughter cells causes the subsequent generation of daughter cells to receive an irregular number of chromosomes.



#### 100 YEARS AGO

The two great Antarctic expeditions have made a stride towards completeness by the launch at Dundee and Kiel of the exploring ships *Discovery* and *Gauss*, both vessels built, at great expense, specially for service in the Antarctic ice. No complete official announcement of the organisation and programme of either expedition has yet been made. However, the two ships are afloat and appear to be the finest vessels for ice-navigation ever constructed, not even excepting the *Fram*, which of course was planned for drifting with the ice-floes, not for sailing through them.

From *Nature* 18 April 1901.

#### 50 YEARS AGO

In the past, the discovery of any man-like fossil has signalled the beginning of extended and heated controversy. After reading numerous recent publications on the South African 'man-apes', we have come to feel that history is repeating itself. The great importance of this remarkable series of fossils is being obscured by technical argument over minor points and by the creation of genera, and even subfamilies, of doubtful utility. The present controversies over the South African fossils started before the discovery of many of the best preserved and most critical specimens, and a general statement of opinion now may help to orientate discussion when these latest finds are described in detail...The human ilium is a very distinctive, highly specialized bone, which is an essential part of the complex that makes possible efficient bipedal locomotion. Among all mammals, the human ilium is approximated only in South African 'man-apes'. If the pelvis only of these forms had been discovered they would surely have been attributed to man, and possibly even to the genus *Homo*. But the skulls associated with these pelvises have small cranial capacities. Thus, the great evolutionary importance of the 'man-apes' is that they show the initial division between ape and man to be a locomotor adaptation... So far as taxonomy is concerned, if the basic adaptation separating the human line from the ape line is bipedal locomotion, then the easiest way to express the facts is to place the 'man-apes' in the family Hominidae. The family might conveniently be divided into two genera, *Australopithecus* and *Homo*. But the main evolutionary significance of the 'man-apes' is independent of the names ultimately applied.

From *Nature* 21 April 1951.



Stelzmann-Strasse 52, 50931 Köln, Germany.  
e-mail: martin.scheffner@uni-koeln.de  
Noel J. Whitaker is in the School of Biochemistry  
and Molecular Genetics, University of New South  
Wales, Sydney 2052, Australia.  
e-mail: n.whitaker@unsw.edu.au

1. Rao, H., Uhlmann, F., Nasmyth, K. & Varshavsky, A. *Nature* **410**, 955–959 (2001).
2. Bachmair, A., Finley, D. & Varshavsky, A. *Science* **234**, 179–186 (1986).
3. Bachmair, A. & Varshavsky, A. *Cell* **56**, 1019–1032 (1989).
4. Hershko, A., Ciechanover, A. & Varshavsky, A. *Nature Med.* **6**, 1073–1081 (2000).
5. Bartel, B., Wunning, I. & Varshavsky, A. *EMBO J.* **9**, 3179–3189 (1990).

6. Kwon, Y. T. *et al. Proc. Natl Acad. Sci. USA* **95**, 7898–7903 (1998).
7. Varshavsky, A., Turner, G., Du, F. & Xie, Y. *Biol. Chem.* **381**, 779–789 (2000).
8. de Groot, R. J., Rumenapf, T., Kuhn, R. J., Strauss, E. G. & Strauss, J. H. *Proc. Natl Acad. Sci. USA* **88**, 8967–8971 (1991).
9. Sijts, A. J., Pilip, I. & Pamer, E. G. *J. Biol. Chem.* **272**, 19261–19268 (1997).
10. Mulder, L. C. & Muesing, M. A. *J. Biol. Chem.* **275**, 29749–29753 (2000).
11. Nasmyth, K., Peters, J. M. & Uhlmann, F. *Science* **288**, 1379–1385 (2000).
12. Uhlmann, F., Wernic, D., Poupart, M. A., Koonin, E. V. & Nasmyth, K. *Cell* **103**, 375–386 (2000).
13. Waizenegger, I. C., Hauf, S., Meinke, A. & Peters, J. M. *Cell* **103**, 399–410 (2000).
14. Cohen-Fix, O., Peters, J. M., Kirschner, M. W. & Koshland, D. *Genes Dev.* **10**, 3081–3093 (1996).

## Materials science

# Polymers all in a row

Galen D. Stucky

Optical and electronic devices made from conducting polymers would be cheaper than silicon technology. Using hybrid materials, such as polymer and silica composites, could lead to high-performance devices.

Plastics that conduct electricity have been around since the 1970s, but their electronic properties, and widespread use, have been limited by structural disorder. The most promising materials are conjugated organic polymers — molecules with a 'backbone' of alternating double and single bonds along which electrons can flow. The mechanical flexibility and tunable optical properties of some conducting polymers make them attractive materials for new optical and electronic devices<sup>1,2</sup>, such as organic light-emitting displays and biomolecular sensors.

The one-dimensional nature of these polymers can be misleading. Thin films of conjugated polymers still have disordered structures, with twists and bends in individual polymer strands. For most applications requiring bulk materials, the challenge is to align the polymer chains uniformly to control the efficiency with which light and electrons can be transported by the polymer molecules. One way to do this is to incorporate conjugated organic polymers into highly structured inorganic materials, such as silica. An inorganic framework provides greater control over the alignment, stability and electronic properties of the polymer chains<sup>3,4</sup>. But existing fabrication processes can be slow and may still yield disordered structures. On page 913 of this issue<sup>5</sup>, Lu *et al.* report a precise and efficient way of synthesizing organic polymers within an inorganic framework.

Previous routes to making organic/inorganic composites have been based on host-guest chemistry, in which a porous 'host' is designed to accept 'guest' molecules. Since their discovery ten years ago<sup>6,7</sup>, silica structures with mesoscale pores (diameters of 2 to 50 nm) have been used to create a number of composite materials by incorporating different guest molecules. For example,

a silica host can be immersed in a conjugated polymer solution so that monomers<sup>4</sup> and polymers<sup>3</sup> infiltrate the empty pores. For the latter case, Tolbert and co-workers<sup>3</sup> have shown that individual conducting polymers can be encased within the pores of mesoporous silica. But such infiltration methods suffer from uneven polymer distributions — besides the well-aligned polymer chains inside the channels, up to 20% of the polymers remain outside the pores<sup>3</sup>, reducing electron and optical energy transfer.

Lu *et al.*<sup>5</sup> demonstrate a new approach to making thin films of silica containing conjugated polymers in a 'one pot' synthesis. Instead of infusing the preformed silica with a polymer solution, they combine organic and inorganic compounds directly by spin-coating solutions containing silicic acid and amphiphilic molecules onto a substrate. Cooperative interactions between the organic and inorganic species create ordered structures on the substrate surface. The amphiphilic molecules have two parts: a hydrophilic and hydrophobic portion, containing an unsaturated hydrocarbon, which can subsequently be transformed into the desired polymer (Fig. 1a). After spin coating, the amphiphilic molecules are homogeneously dispersed within the silica matrix and the unsaturated hydrocarbon molecules can be turned into conjugated polymers with a polydiacetylene backbone (PDA) on exposure to ultraviolet radiation. By varying the length of the hydrophilic part of the amphiphilic molecule, highly ordered thin films with aligned pores in layered, hexagonal and cubic patterns can be formed (Fig. 1b). The polymer/silica composite has pores of diameter 2.5 to 3 nm — just about large enough for a single PDA chain with its coordinated hydrophilic groups.

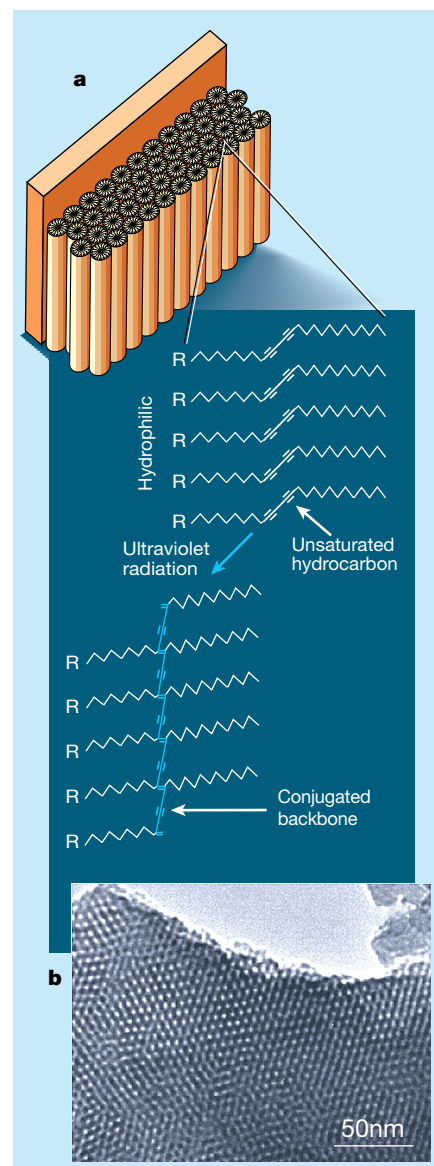


Figure 1 The polymer/silica composite created by Lu *et al.*<sup>5</sup>. a, After some initial processing, the channels of the silicic acid framework contain aligned amphiphilic molecules, each of which ends in a hydrophilic R group (polyethyleneoxide chain). After ultraviolet irradiation, the molecule polymerizes (the conjugated backbone is shown in blue). b, A transmission electron micrograph of the final polymer/silica composite, showing the cubic patterns of the silica channels.

One benefit of making structures that tightly control the packaging and orientation of the polymer molecules is that they have enhanced optical properties. Conjugated PDA polymers can absorb light over two distinct frequency ranges: the red band (absorption maximum at about 540 nm) and the blue band (absorption maximum at about 645 nm). The change from blue to red is caused by a shortening of the effective conjugation length as a result of changes in the conformation of the polymer backbone. The effective conjugation length depends on the loops and bends in the polymer molecule — it is related

to the distance along which the electrons can interact on the backbone of the polymer.

In Lu and his colleagues' study<sup>5</sup>, illuminating the composite with ultraviolet radiation causes polymerization of the monomer and results in the blue form. The blue colour of the synthesized films indicates that the monomers are highly polymerized, because the chains are well aligned, resulting in a longer conjugation length. The surrounding silica structure not only orients the polymer chains, but also stabilizes the PDA polymers and protects them from oxidation by the air.

In addition, Lu *et al.* show that their polymer/silica films can be switched from blue to red by changes in temperature, solvent or mechanical properties, as expected for highly ordered PDA. When blue PDA/silica films are heated above 47 °C, a rapidly reversible change of colour is observed. The fast reversibility is explained by chemical restoring forces due to hydrogen bonding of the hydrophilic side chains of the polymer with the inorganic silica framework. When the PDA/silica composites are immersed in different solvents, their colour response can be measured from changes in their absorbance at 645 nm. It was found that the absorbance changes linearly with the dielectric constant of the solvent — a measure of its polarity or

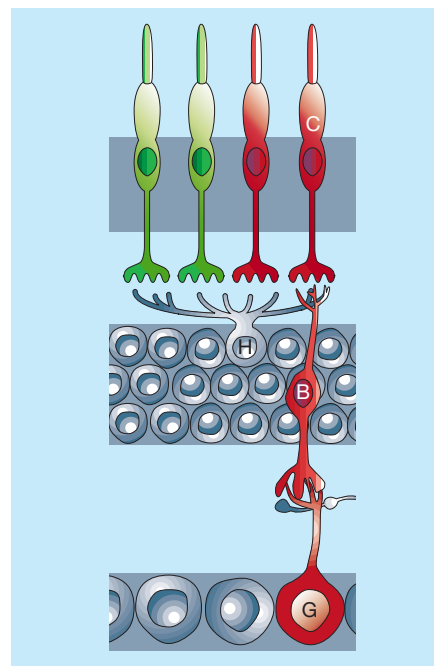
charge structure. The changes in absorbance occur because polar solvents can diffuse more readily within the polar side chains of the polymer. The resulting solvation stresses then lead to a reduction of the effective conjugation length in the polymer backbone. Finally, mechanical abrasion of the blue PDA/silica film causes it to change to the red form, which suggests that new sensor devices can be developed that respond to mechanical damage.

Lu *et al.*<sup>5</sup> have demonstrated a powerful and clean way of producing conjugated, aligned and mechanically robust polymer-based optoelectronic materials. Possible applications of such high-performance composites range from sensors to light-emitting diodes and lasing materials that will be cheap and quick to produce. ■

Galen D. Stucky is in the Materials Department and the Department of Chemistry and Biochemistry, University of California, Santa Barbara, California 93106-9510, USA.

e-mail: stucky@chem.ucsb.edu

1. McGehee, M. D. & Heeger, A. J. *Adv. Mater.* **12**, 1655–1668 (2000).
2. Tessler, N. *Adv. Mater.* **11**, 363–370 (1999).
3. Nguyen, T.-Q. *et al. Science* **288**, 652–656 (2000).
4. Wu, C. G. & Bein, T. *Science* **264**, 1757–1759 (1994).
5. Lu, Y. *et al. Nature* **410**, 913–917 (2001).
6. Yanagisawa, T. *et al. Bull. Chem. Soc. Jpn* **63**, 988–992 (1990).
7. Kresge, C. T., Leonowicz, M. E., Roth, W. J., Vartuli, J. C. & Beck, J. S. *Nature* **359**, 710–712 (1992).



**Figure 1 Photoreceptor wiring for a midget ganglion cell in the central retina.** In the centre of the receptive field, a single cone (C) connects to the ganglion cell (G) through a midget bipolar cell (B). In the surround, several cones connect through a horizontal cell (H) and the bipolar cell. In the intermediate peripheral retina (not shown), several bipolar cells converge on each ganglion cell, and in far periphery several cones connect directly to each bipolar cell. (Redrawn from Fig. 3A of ref. 7.)

## Vision

# Why do colours fade at the edges?

Andrew Derrington

Colour vision is much poorer in peripheral parts of the retina than in the centre. It seems that the usual explanation for that finding is flawed.

In humans and Old World monkeys, good discrimination between colours requires appropriate connections from the different types of photoreceptor in the eye. The process was thought to be compromised in the peripheral parts of the retina. As Martin *et al.* show on page 933 of this issue<sup>1</sup>, however, that is not the case — a finding that will make uncomfortable reading for visual psychophysicists.

There are three classes of photoreceptor — the L, M and S cones, so-called because they are maximally sensitive to long, middle or short wavelengths of light. It is the first two that concern us here. Awareness of the relative activations of the different types of cone depends on their differential connection to cells in the retina, known as ganglion cells, that transmit visual information from the eye to the brain. Forming the appropriate connections appears to be accomplished easily in the cone-rich central retina — the part that we use to look straight at an object. In the periphery of the retina, which deals with the surroundings of what we look at, it is much more difficult because many

photoreceptors converge on each retinal ganglion cell.

Martin *et al.*<sup>1</sup> describe retinal ganglion cells in the periphery of the macaque monkey retina that, surprisingly, have differential inputs from the L and M cones. The finding raises two questions. The first is how the cones become wired up differentially (the authors have an answer for this, which depends on irregularities in the shape of the receptive field). The second question is more subtle, but raises some uncomfortable prospects. The ability of humans to distinguish colours deteriorates towards the periphery of the retina. So, if the ganglion cells are wired up correctly to preserve colour signals, why is that so?

The receptive field of a retinal ganglion cell is the part of the visual field from which light influences the firing rate of the cell. The receptive fields studied by Martin *et al.*, known as type I (ref. 2), are divided into two concentric zones, centre and surround, within which light has opposite effects. If light in the centre excites the cell (increases its firing rate), light in the surround will inhibit it, and vice versa.

The differential or 'colour opponent' photoreceptor input to type I receptive fields takes the form that the centre and surround are driven by different cone classes (Fig. 2). So if light covers the entire receptive field, the cell signals the difference between the activation of the cone type that drives the centre and the cone type that drives the surround.

In the central retina, differential wiring of photoreceptor classes to ganglion cells is easily achieved. The centre of the receptive field contains only a single cone, which connects directly through a midget bipolar cell to the ganglion cell<sup>3</sup> (Fig. 1). Even though the surround contains a mixture of cone classes, connected indirectly through horizontal cells and the bipolar cell to the ganglion cell, the cell still generates a good colour opponent signal<sup>4</sup> (Fig. 3a). In the far peripheral retina, the part beyond 50° from the centre, colour opponency is impossible because several cones converge directly on bipolar cells<sup>5</sup>. However, there is an intermediate zone where midget bipolar cells still receive direct inputs from single cones, but several midget bipolar cells converge on a single ganglion cell.

Unless the midget ganglion cells in this intermediate zone receive differential inputs from bipolar cells driven by different classes of cone, they will not be colour opponent. In this intermediate zone, human sensitivity



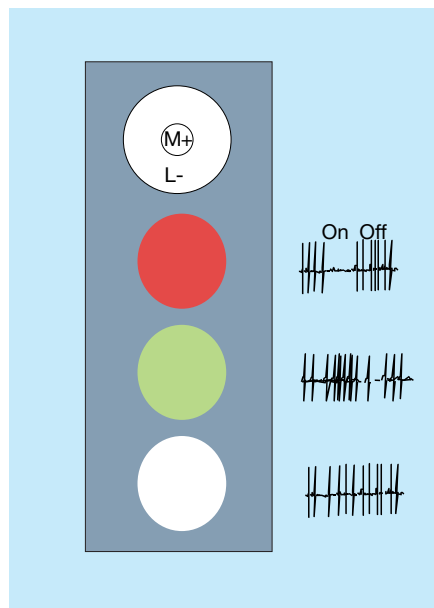


Figure 2 Colour opponency and a typical type I receptive field. Here, light of medium wavelength impinging on the centre of the receptive field excites the cell (M+), whereas long-wavelength input to the surround inhibits it (L-). These colour-opponent signals are essential to colour perception. In this case the responses (the traces on the right) show that the cell would be excited by green light, inhibited by red light and relatively unaffected by white light.

to colour falls off in a way that is consistent with a random wiring between bipolar cells and retinal ganglion cells<sup>6</sup>. So it has been assumed that the midget ganglion cells fail to preserve the colour opponency of their bipolar cell inputs. But when Martin *et al.*<sup>1</sup> directly measured the colour opponency of midget ganglion cells they found that more than half of those in the intermediate zone of peripheral retina (lying 20° to 50° from the central retina) are just as strongly colour opponent as those in the centre.

Martin *et al.* argue that the elliptical shape of the receptive-field centres in midget ganglion cells (they have an aspect ratio of about 1.8) is probably related to their colour selectivity. They show that a ganglion cell that receives input from all the bipolar cells in an appropriately oriented elliptical area can be highly colour selective (Fig. 3b). So the elliptical receptive-field centres could be the product of an adaptive developmental process that has selected bipolar cells driven by the same class of photoreceptor for each ganglion cell. One way to test this would be to examine ganglion cells in monkeys that had been reared in monochromatic light, which would block any selection process based on differential responses to colours.

Martin and colleagues' result<sup>1</sup> is neat — and embarrassing. The discrepancy between the physiological result and human psychophysics cannot be ignored because Martin *et al.* confirmed the psychophysical result in

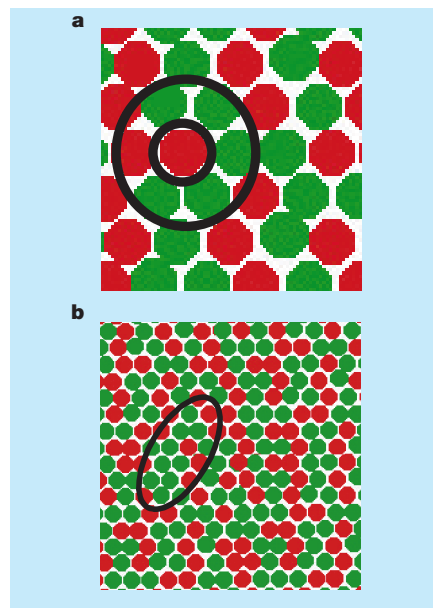


Figure 3 Colour opponency in the centre and periphery of the retina. a, A midget ganglion cell whose centre is driven by a single cone, as happens in the central retina, is almost guaranteed to have colour opponency. b, Martin *et al.*<sup>1</sup> propose that the same effect can be achieved in the near periphery (20° to 50° from the centre) if the centre of the receptive field is elliptical and oriented so that most of its input comes from a single class of cone.

humans using the same stimuli that they had used for the physiology. There are two possible ways out, neither of which is very palatable. One is that, during processing of visual signals in the brain, there is a failure to use the information provided by colour-opponent ganglion cells in the near periphery of the retina. But this would introduce a problem for psychophysicists because they infer the properties of the visual system from measurements of visual performance. The other is that monkey and human retinas may be different, which again would be a tough conclusion to reach because the monkey has always been taken as a close physiological model for human vision. These doubts can be resolved only with data — either from psychophysical measurements on the periphery of monkey retinas, or from physiological work on the human retina. ■

Andrew Derrington is in the School of Psychology, University of Nottingham, University Park, Nottingham NG7 2RD, UK.

e-mail: andrew.derrington@nottingham.ac.uk

1. Martin, P. R., Lee, B. B., White, A. J. R., Solomon, S. G. & Rüttiger, L. *Nature* **410**, 933–936 (2001).
2. Wiesel, T. N. & Hubel, D. H. *J. Neurophysiol.* **29**, 1115–1156 (1966).
3. Boycott, B. B. & Dowling, J. E. *Phil. Trans. R. Soc. Lond. B* **225**, 109–184 (1969).
4. Lennie, P., Haake, P. W. & Williams, D. R. in *Computational Models of Visual Processing* (eds Landy, M. S. & Movshon, J. A.) 71–82 (MIT Press, Cambridge, MA, 1991).
5. Wässle, H. *et al. Vision Res.* **34**, 561–579 (1994).
6. Mullen, K. T. & Kingdom, F. A. A. *Vision Res.* **36**, 1995–2000 (1996).
7. Calkins, D. J. & Sterling, P. *Neuron* **24**, 313–321 (1999).

## Daedalus

### Adhesive vapours

Paper, felt and many other useful materials are made by wetting a mass of fine fibres or particles and letting them dry out. As they do so, surface tension forces the particles together. They adhere strongly when they are dry. Sadly, they tend to disintegrate again in the wet.

Curiously though, even dry particles can stick together by aqueous surface tension. Glass beads, for example, adhere weakly but spontaneously by the capillary condensation of a tiny amount of atmospheric humidity at their points of contact. Daedalus sees this as a route to a whole new range of aggregated materials. He recalls that methyl cyanoacrylate (superglue) is volatile, but polymerizes only in the liquid state. So he is exposing compact masses of cellulose fibres or plaster crystals to an atmosphere not quite saturated with cyanoacrylate vapour. It will condense by capillarity only at their points of contact. The liquid will then set, aggregating the particles with the absolute minimum of glue, applied exactly where it is needed. The resulting solid will be far stronger than ordinary plaster or paper, and will be waterproof.

Almost any combination of powders or fibres could be aggregated in this way. New composites of chalk, sawdust, soot, cotton fibres, iron filings and straw, or mixtures of these, should be possible, combining cheapness with many unexpected virtues. Non-woven fabrics, paper made from silica fibres or fine wire strands, porous diatomaceous concrete, wonderfully light synthetic eiderdown fluff, all seem possible.

But Daedalus wants to extend the idea to metals, as a sort of vapour-phase soldering. Imagine, he says, an electrical circuit unit with all the leads in place, touching the right components or contacts but not soldered to them. Expose the whole thing to unsaturated metal vapour: lead, tin or even mercury. Capillary condensation would condense it exactly at all the points of contact, making all the connections perfectly. Mercury would rapidly amalgamate with the metal it had condensed on, giving a stable non-volatile joint.

At present, microchips are often wired to their carriers by human operators using microscopes and micromanipulators. Vapour-phase capillary-condensation soldering could do the job automatically, at great speed, and with no need for precise manipulation. And nanotechnology would have a whole new method of microfabrication.

David Jones

Obituary

# Minoru Oda (1923–2001)

Leaders of research in Japan come in two contrasting styles. There are martinets in the mould of the traditional university professor, who issue orders and expect their subordinates to carry them out. And there are cultivated men (but not many women yet) whose method is to encourage lesser mortals to work up their ideas into projects and then give them wholehearted support. Minoru Oda, who died on 1 March 2001, was emphatically in the second style. He was also recognized as one of the dozen or so people who had largely modernized Japan's science in the quarter of a century beginning around 1965, at least partially wresting management (and even micro-management) of basic research from feuding government departments.

Oda was born in rural Sapporo in 1923 and won a place at Osaka University (then the imperial university of the same name). He escaped war service, graduating in 1944, and immediately embarked on graduate study of cosmic rays. His destiny seemed to be a traditional academic career, leading eventually to a martinetship. But that pattern was broken by a three-year spell at the Massachusetts Institute of Technology (MIT), beginning in 1953, which he later said had changed his life. He learned that research, meritocratic though it must be, could be a democratic process too. He also used his pre-Sputnik time at MIT to cultivate an enduring interest in X-ray astronomy.

He returned not to Osaka but to Tokyo, where he became one of the ring-leaders of the project hatched within the physics department to build and launch Earth-orbiting satellites. Elsewhere — even at MIT — the idea that university departments might nurse such an ambition might seem to be a joke: at Tokyo, interministerial rivalry helped to make it a reality. At a time when the government was planning to set up a national space programme to be carried out by a new organization (the National Space Development Agency as it has become), shrewd observers could easily calculate that the education ministry (then known as Monbusho) might fall for a plea to launch a cheaper and quicker scheme under its own control. That was the origin of the Institute of Space and Aeronautical Science (ISAS), of which *Nature* asked, a few years ago, "Is this the best lab in the world?"

After a further stint at MIT, Oda returned as project director at the space



## Guiding hand in Japanese science in the late twentieth century

institute, which in due course became a national facility to which all Japanese universities have access. He became director-general of the organization in 1984, which in Japanese custom is a pre-retirement post, and became chairman of its council a few years later. The institute's virtues are that it has been able to launch Earth satellites cheaply and without fuss. Several of these have broken new ground in, for instance, the use of soft X-rays to study the Sun's surface, and in using pairs of satellites for making long-baseline interferometry measurements at radio and microwave frequencies.

ISAS may be Oda's most conspicuous and tangible legacy, but it is easy to underestimate his accomplishment. Cannot anybody set up a new lab, if only the government will hand out the funds? In reality, success in a daring enterprise such as this rests on the accuracy of the judgement that success is possible, for otherwise the money-bags will be closed. Persuading them to open in the first place requires not merely charm (of which Oda had plenty) but guile and determination. His achievement at ISAS is a proof of his confidence in his colleagues and of theirs in him. His confidence in himself was habitually belied by his soft-spoken modesty.

The same talents helped Oda to make a success of his spell from 1988 to 1993 as president of one of Japan's other iconoclastic laboratories, RIKEN. This is a general-purpose laboratory, founded in 1917 with a benefaction from the then Emperor, when it first dawned on some Japanese that they had failed to keep up with the rest of the world in science. The laboratory is now a dependant of the education ministry, and has grown hugely in size to several thousand people, but it retains a large degree of independence. Although most of its work is in physical science and engineering, a few years ago it was able to jump into work on the Human Genome Project. Oda was a staunch defender of RIKEN's freedom to do what it believed to be important, provided that it could see a way of succeeding. During this period, he also worked hard at increasing the numbers of foreign scientists at RIKEN.

For his final essay in institution-building, in 1994 Oda became director of the newly established International Institute for Advanced Studies near Osaka. This again was an opportunity for casting a net beyond Japan. Oda had a particular fondness both for the United States and for France (where his daughter Reiko now works).

Ironically, Oda found that the last few months of his life were occupied with anxiety that the impending reorganization of Japanese science may undermine some of the achievements of recent decades. The particular proposal that the two national space agencies should be merged risks the ethos of what began as the University of Tokyo's project. Hierarchical decision-making, Oda's anathema, is the obvious danger.

Luckily, the mood in Japan has radically changed. The education ministry has learned from the likes of Oda that research is a delicate process requiring the consent and enthusiasm of those who carry it out; the ministry may yet find ways of carrying through policies that reflect its new knowledge that not all universities are equal. And there is now a small army of people in universities and research institutes holding the views about the conduct of research that Oda's associates have championed. Probably there is no turning back.

Oda is survived by his two children, both scientists, and by his wife, Tomo Oda, a woman of outstanding elegance even by Japanese standards. **John Maddox**  
*John Maddox is emeritus editor of Nature.*



# Subsidence risk from thawing permafrost

The threat to man-made structures across regions in the far north can be monitored.

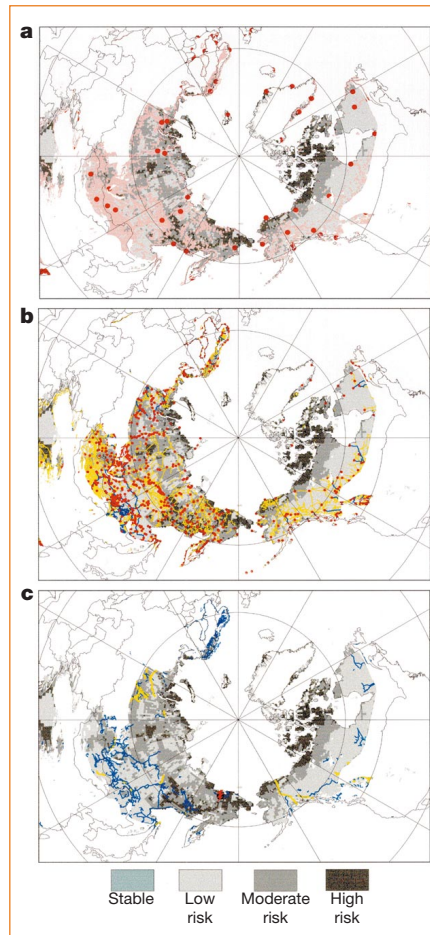
The thawing and disappearance of permafrost has accelerated in recent decades<sup>1</sup>, damaging buildings and infrastructure and causing public concern<sup>2</sup>. Here we offer a geographic overview of the hazard potential associated with thawing permafrost in the Northern Hemisphere which indicates that vulnerability to subsidence is widespread. Much of the existing infrastructure erected in northern regions is located in areas of high hazard potential and could be affected by thaw subsidence under conditions of global warming.

Climate-change scenarios predict a marked warming at high latitudes<sup>3</sup>. Instrument records indicate that warming is underway in much of the Arctic, affecting sea ice, glaciers, vegetation and the duration of lake ice<sup>1</sup>. Thawing of ice-rich permafrost can lead to subsidence and deformation of level surfaces into irregular 'thermokarst' terrain. Effects range from localized subsidence beneath individual structures, to deep and extensive depressions developed in response to long-term climate change.

Thawing of permafrost layers near the surface can have a severe impact on engineered structures. On slopes, the process creates mechanical discontinuities in the substrate, leading to potentially destructive slides, flows and slumps. Wave action along ice-rich shorelines can produce very high rates of erosion.

Cold permafrost (in the continuous zone) and relatively warm permafrost (discontinuous and sporadic zones) differ in their response to climatic warming. The primary effects of warming on cold permafrost are increased temperature and thickening of the seasonally thawed layer, but may be sufficient to affect engineered structures. The reaction of warm permafrost to temperature increases can be more pronounced. Taliks (unfrozen ground) develop between the layer of seasonal freezing and the permafrost table, and permafrost may disappear altogether. Depending on ice content, settlement of several metres or more is possible.

We used the general circulation model (GCM) climate scenario known as GFDL89 (ref. 4), a mathematical solution for the thickness of the seasonally thawed layer<sup>5</sup>, a digital representation of soil properties<sup>6</sup>, and permafrost distribution and ice content<sup>7</sup> to construct hazard-zone maps<sup>8</sup> for the Northern Hemisphere. Of three GCM scenarios designated by the Intergovernmental Panel on Climate Change (IPCC)<sup>4</sup> and used to study potential changes in permafrost distribution<sup>9</sup>, the GFDL89 model provides mid-



**Figure 1** Permafrost hazard potential in the Northern Hemisphere. Locations of existing settlements and infrastructure<sup>12</sup> are shown: **a**, population centres (red), settlements (pink); **b**, undifferentiated roads and trails (yellow), railroads (blue), airfields (red); **c**, electrical transmission lines (blue), pipelines (yellow); Bilibino nuclear station (red). Hazard-potential zonation is based on a settlement index, given by  $I_s = \Delta z_s \times V_{ice}$ , where  $\Delta z_s$  is the relative change in maximum seasonal thaw depth under the GFDL89 GCM scenario<sup>5</sup> and  $V_{ice}$  is the volumetric proportion of near-surface soil occupied by ground ice<sup>7</sup>. Class intervals for hazard-zone categories were developed by applying a nested-means procedure<sup>13</sup> to the pooled frequency distribution of  $I_s$  values from three IPCC climate scenarios<sup>4</sup>. We assume drainage of liquid water, and that subsidence is proportional to volume of ice lost. Data layers are registered at a resolution of 0.5° latitude/longitude. Further details are available from the authors.

range results for hazard potential. A dimensionless 'settlement index' was used to classify contemporary permafrost according to its vulnerability to thaw subsidence under global warming.

Areas at risk can be partitioned (Fig. 1) into those with low, moderate or high susceptibility to thaw-induced settlement. A zone in the high-risk category extends discontinuously around the Arctic Ocean,

indicating that there is a high potential for coastal erosion. Within this region are population centres (Barrow, Inuvik) and river terminals on the Arctic coast of Russia (Salekhard, Igarka, Dudinka, Tiksi). Transportation and pipeline corridors traverse areas of high hazard potential in northwestern North America.

The area containing the Nadym–Pur–Taz natural-gas production complex and its associated infrastructure in northwest Siberia is also at high risk (Fig. 1). Large areas elsewhere in Siberia, particularly Sakha–Yakutia and the far east, show moderate or high hazard potential. Within these areas are several large population centres (Yakutsk, Noril'sk, Vorkuta), an extensive network of roads and trails, and the trans-Siberian and Baikal–Amur mainline railroads. The Bilibino nuclear power station and its grid occupy an area at high risk in the Russian far east. Areas at lower risk are associated with mountainous terrain and stable cratons in which bedrock is at or near the surface.

Detrimental changes in permafrost conditions may evolve gradually<sup>10</sup> and can be predicted and monitored, allowing mitigation of negative consequences and avoidance of catastrophic events. The generalized delineation in Fig. 1 shows areas to which high priority should be assigned for monitoring permafrost conditions, and could also form a basis for planning and developing assessments at other scales<sup>11</sup>.

**Frederick E. Nelson\***, **Oleg A. Anisimov†**, **Nikolay I. Shiklomanov\***

*\*Department of Geography and Center for Climatic Research, University of Delaware, Newark, Delaware 19716, USA*

*e-mail: fnelson@udel.edu*

*†State Hydrological Institute, Second Line VO 23, 199053 St Petersburg, Russia*

1. Serreze, M. C. et al. *Climatic Change* **46**, 159–207 (2000).
2. Linden, E. *Time* **156**, 52–56 (2000).
3. Kattenberg, A. et al. in *Climate Change 1995: The Science of Climate Change. Contribution of Working Group I to the Second Assessment Report of the Intergovernmental Panel on Climate Change* (eds Houghton, J. J. et al.) 285–357 (Cambridge Univ. Press, New York, 1996).
4. Greco, S. et al. & IPCC Working Group II. *Climate Scenarios and Socioeconomic Projections for IPCC WG II Assessment* (Consortium for International Earth Science Information Network, Washington DC, 1994).
5. Anisimov, O. A., Shiklomanov, N. I. & Nelson, F. E. *Glob. Planet. Change* **15**, 61–77 (1997).
6. Staub, B. & Rosenzweig, C. *Global Gridded Data Sets of Soil Type, Soil Texture, Surface Slope and Other Properties* (National Center for Atmospheric Research, Boulder, CO, 1987). URL: <http://www.ngdc.noaa.gov/seg/eco/cdroms/gedii/datasets/a11/sr.htm#top>
7. Brown, J., Ferrians, O. J. J., Heginbottom, J. A. & Melnikov, E. S. *International Permafrost Association Circum-Arctic Map of Permafrost and Ground Ice Conditions* (scale 1:10,000,000, US Geological Survey Circum-Pacific Map Series, Map CP-45, Reston, VA, 1997). URL: <http://www.geodata.soton.ac.uk/ipa/>

8. Monmonier, M. *Cartographies of Danger: Mapping Hazards in America* (Univ. Chicago Press, 1997).
9. Anisimov, O. A. & Nelson, F. E. *Climatic Change* **35**, 241–258 (1997).
10. Riseborough, D. W. in *Proc. Fifth Canadian Permafrost Conf.* (eds Burgess, M. M., Harry, D. G. & Sego, D. C.) 199–205 (Centre d'études nordiques, Université Laval/National

- Research Council of Canada, Québec, 1990).
11. Parmuzin, S. Y. & Shamanova, I. I. *Polar Geogr. Geol.* **10**, 184–193 (1986).
12. US National Imagery and Mapping Agency. *Digital Chart of the World for ADOL* (ESRI, Inc., Redlands, CA, 1997). URL: <http://www.esri.com/data/online/esri/wobmsselect.html>
13. Scripser, M. W. *Ann. Assoc. Am. Geogr.* **60**, 385–393 (1970).

## Neuroperception

# Early visual experience and face processing

Adult-like expertise in processing face information takes years to develop<sup>1</sup> and is mediated in part by specialized cortical mechanisms<sup>2</sup> sensitive to the spacing of facial features (configural processing)<sup>3</sup>. Here we show that deprivation of patterned visual input from birth until 2–6 months of age results in permanent deficits in configural face processing. Even after more than nine years' recovery, patients treated for bilateral congenital cataracts were severely impaired at differentiating faces that differed only in the spacing of their features, but were normal in distinguishing those varying only in the shape of individual features. These findings indicate that early visual input is necessary for normal development of the neural architecture that will later specialize for configural processing of faces.

For face recognition, subtle differences in the shape of specific features (featural information) and/or in their spacing (configural information) must be encoded. We created two sets of faces that differentiated configural from featural processing<sup>4</sup> (Fig. 1a, b). We asked 26 normal right-handed adults to view the faces from each set binocularly to decide whether they were the same or different when upright and inverted. Adults were equally accurate in differentiating faces from the two sets in their canonical upright position (Table 1). Inverting the faces decreased adults' accuracy for the configural but not the featural set, consistent with previous findings<sup>4,5</sup>.

We used these stimuli to test face processing in 14 patients (6 male; 13 right-handed; 11 caucasian) born with a dense central cataract in each eye that prevented patterned stimulation from reaching the retina<sup>6</sup>. After removal of the natural lens, an optical correction was fitted to focus visual input (mean duration of deprivation, 118

days from birth; range, 62–187 days); patients had had at least nine years of visual experience after treatment before testing and, when necessary, wore an additional optical correction to focus the eyes at the testing distance.

Compared with age-matched normal subjects, the deprived patients distinguished faces from the featural set normally, but were significantly impaired for the configural set of upright faces (Table 1). Performance was not related to the duration of the deprivation for either set ( $P > 0.10$ ; Fig. 1c). There was no correlation between acuity (range from 20/25 to 20/80 in the better eye; median, 20/40) and patients' accuracy on either set ( $P > 0.10$ ).

Our results indicate that visual experience during the first few months of life is necessary for the normal development of expert face processing. Because normal infants have poor visual acuity<sup>7</sup>, their cortex is exposed only to information of low spatial frequency which, for faces, specifies the global contour and location of features but little of their detail<sup>8</sup>. This early information sets up the neural architecture that will specialize in expert configural processing of faces over the next 10–12 years<sup>9,10</sup>. When visual input is delayed by as little as two months, permanent deficits result.

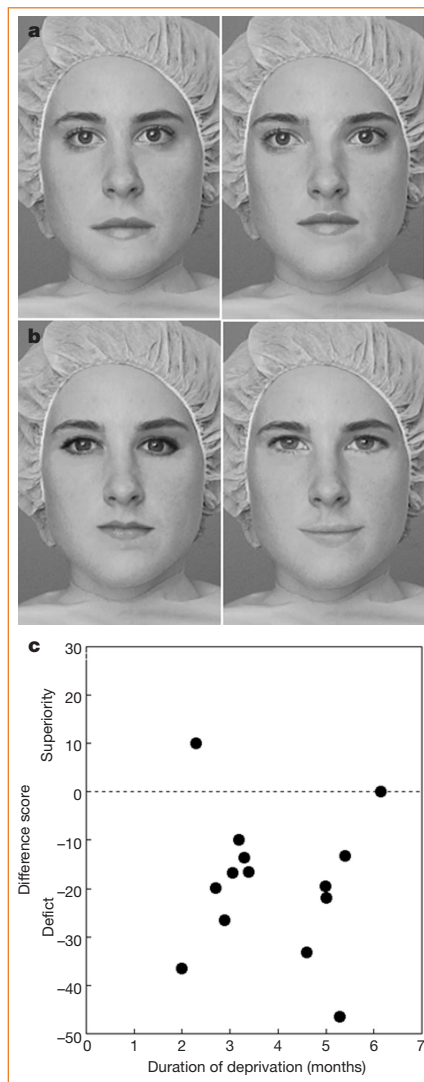
Patients performed normally in a different task requiring the discrimination of geometric patterns based on the location of an internal feature<sup>11</sup>. This suggests that deficits in configural processing may be restricted to the processing of faces, as expected from evidence that normal adults use separate systems for processing face and non-face objects<sup>12,13</sup>.

**Richard Le Grand\***, **Catherine J. Mondloch\***, **Daphne Maurer†**, **Henry P. Brent†**

\*Department of Psychology, McMaster University, Hamilton L8S 4K1, Canada  
e-mail: [maurer@mcmaster.ca](mailto:maurer@mcmaster.ca)

†The Hospital for Sick Children, 555 University Avenue, Toronto M5G 1X8, Canada

1. Carey, S. & Diamond, R. *Vis. Cognit.* **1**, 253–274 (1994).



**Figure 1** Measurement of configural versus featural processing. Examples of stimuli from **a**, the configural set (created by moving the eyes and mouth), and **b**, the featural set (created by replacing the eyes and mouth). On each trial, one of the five possible faces appeared for 200 ms, and following an interstimulus interval of 300 ms, a second face appeared until the subject used a joystick to signal a 'same' or 'different' judgement. **c**, Patients' performance on the upright configural set plotted as a function of the duration of deprivation from birth. Each circle represents the difference between the accuracy (per cent correct) of one patient and his/her age-matched control.

**Table 1** Detection of facial configural and featural differences

| Group    | N  | Mean age (range) | Configural |          | Featural |          |
|----------|----|------------------|------------|----------|----------|----------|
|          |    |                  | Upright    | Inverted | Upright  | Inverted |
| Adults   | 26 | 19 (18–22)       | 80 (1.9)   | 63 (2.2) | 80 (1.4) | 81 (1.8) |
| Controls | 14 | 14 (9–21)        | 81 (2.7)   | 59 (1.8) | 79 (1.9) | 80 (2.9) |
| Patients | 14 | 14 (9–21)        | 62 (3.2)   | 55 (3.0) | 80 (2.4) | 78 (3.1) |

Mean accuracy (per cent correct) and standard error are shown for detecting configural and featural differences in upright and inverted faces. There were 30 trials for each of the four conditions, which were presented in separate blocks. For adults, inversion decreased accuracy only for the configural set (significant interaction between orientation and stimulus set ( $F_{1,25} = 49.78$ ,  $P < 0.01$ ), followed by analysis of simple effects ( $F_{1,27} = 62.57$ ,  $P < 0.01$ )). Patients were less accurate than controls (matched on age, gender, race and handedness) only for the upright configural set (significant interaction between group, orientation and stimulus set ( $F_{1,26} = 8.99$ ,  $P < 0.01$ ), followed by analysis of simple effects ( $F_{1,26} = 16.26$ ,  $P < 0.01$ )).

2. McCarthy, G., Puce, A., Gore, J. & Allison, J. *Cognit. Neurosci.* **9**, 605–610 (1997).
3. Bartlett, J. C. & Searcy, J. *Cognit. Psychol.* **25**, 281–316 (1993).
4. Freire, A., Lee, K. & Symons, L. A. *Perception* **29**, 159–170 (2000).
5. Tanaka, J. W. & Sengco, J. A. *Mem. Cognit.* **25**, 583–592 (1997).
6. Ellemberg, D., Lewis, T. L., Maurer, D., Liu, C. H. & Brent, H. P. *Vis. Res.* **39**, 3480–3489 (1999).
7. Maurer, D. & Lewis, T. L. *Clin. Neurosci. Res.* (in the press).
8. de Schonen, S. & Mathivet, E. *Eur. Bull. Cogn. Psychol.* **9**, 3–44 (1989).
9. Diamond, R. & Carey, S. *J. Exp. Psychol. Gen.* **115**, 107–117 (1986).
10. Carey, S., Diamond, R. & Woods, B. *Dev. Psychol.* **16**, 257–269 (1980).
11. Geldart, S. thesis, McMaster Univ. (2000).
12. Kanwisher, N., McDermott, J. & Chun, M. M. *J. Neurosci.* **17**, 4302–4311 (1997).
13. Farah, M. J. *The Cognitive Neuroscience of Vision* (Blackwell, Malden, 2000).



# Increased sedimentation rates and grain sizes 2–4 Myr ago due to the influence of climate change on erosion rates

Zhang Peizhen\*, Peter Molnar†‡ & William R. Downs§

\*Institute of Geology, State Seismology Bureau, Beijing, China

‡Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

§Bilby Research Center, Northern Arizona University, Flagstaff, Arizona 86011-6013, USA

**Around the globe, and in a variety of settings including active and inactive mountain belts, increases in sedimentation rates as well as in grain sizes of sediments were recorded at ~2–4 Myr ago, implying increased erosion rates. A change in climate represents the only process that is globally synchronous and can potentially account for the widespread increase in erosion and sedimentation, but no single process—like a lowering of sea levels or expanded glaciation—can explain increases in sedimentation in all environments, encompassing continental margins and interiors, and tropical as well as higher latitudes. We suggest that climate affected erosion mainly by the transition from a period of climate stability, in which landscapes had attained equilibrium configurations, to a time of frequent and abrupt changes in temperature, precipitation and vegetation, which prevented fluvial and glacial systems from establishing equilibrium states.**

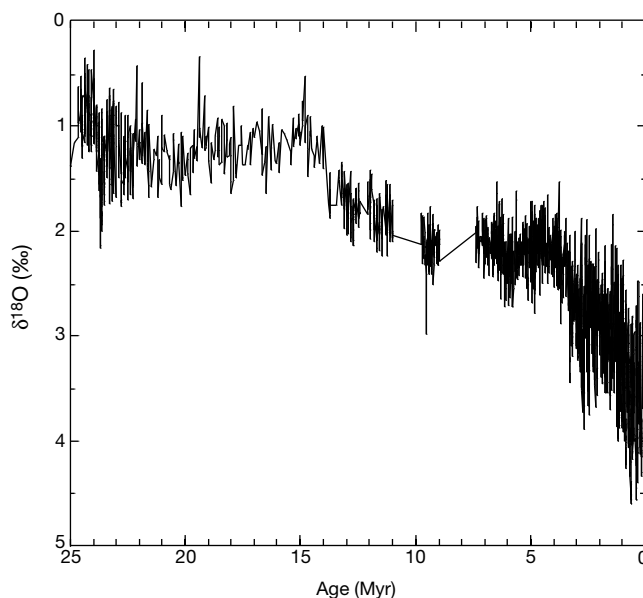
Two general processes govern rates of regional erosion and associated rates of terrigenous sedimentation. (1) Crustal deformation creates regionally elevated terrain that provides potential energy to rivers and glaciers, the main agents of erosion. (2) Climate, which has changed globally since 3–4 Myr ago (Fig. 1; refs 1–3), controls the abundance and temporal distribution of these erosive agents, as well as the type, density and distribution of vegetation that protects the Earth's surface from erosion. Geologists working in isolated mountain belts have repeatedly concluded that the ranges they studied rose in late Pliocene and Quaternary time (since ~3 Myr ago) (see Supplementary Information), buttressing such inferences with observations that sediment has accumulated rapidly near the ranges since ~5 Myr ago.

Sediment accumulation rates throughout the world show increases of two to as much as 10 times beginning 2–4 Myr ago (Fig. 2). Although some mountainous environments surely have grown since that time, such conclusions for mountain belts throughout the world, when taken together, would call for a global geodynamic process to orchestrate such synchronicity. Astronomical calibration of the geomagnetic timescale rules out globally synchronous changes in tectonic-plate motions since 9 Myr ago (ref. 4). Although rates of motion between pairs of plates have changed in this interval, such changes do not involve all plates. Climate change offers the most plausible, globally synchronous process affecting erosion, especially given that both active and inactive mountain belts show accelerated erosion since ~3–4 Myr ago (ref. 5). The question then arises as to what aspect of climate change is responsible for increased erosion. Different aspects such as sea-level change or increased glaciation, could contribute to increased erosion and sedimentation rates in different regions, but none provides an explanation for all regions, and particularly those unaffected by sea level or glaciation. The most important change may not be the switch from one state (warm and humid) to another (cold and dry) at 3–4 Myr ago. Rather, the change from a virtually unchanging climate to one that has been changing rapidly, as dictated by Milankovitch forcing (related to changes in the orbital

parameters of the Earth) and amplified by a periodically ice-covered Earth, might maintain a state in which erosive processes never reach equilibrium with the evolving landscape.

## Global increase in sedimentation

In a compilation of data relating to sediment accumulation in the main oceans, Hay *et al.*<sup>6</sup> showed that terrigenous sedimentation has



**Figure 1** Plot of  $\delta^{18}\text{O}$  from benthic foraminifers since 25 Myr ago, showing increases in mean values and in variability since ~4 Myr ago. The former increases imply cooling, and the latter increases imply an increasingly variable climate. Values (in ‰) have been measured largely (~95%) from fossil tests of *Cibicides* spp., or adjusted to be equivalent to those of *Cibicides* (ref. 63), from the Ceara rise in the eastern equatorial Atlantic Ocean (Ocean Drilling Project sites 925, 926 and 926). Values are plotted increasing downwards to reflect cooling. Data are from refs 62–66, and from T. Bickert and W. B. Curry, personal communication.

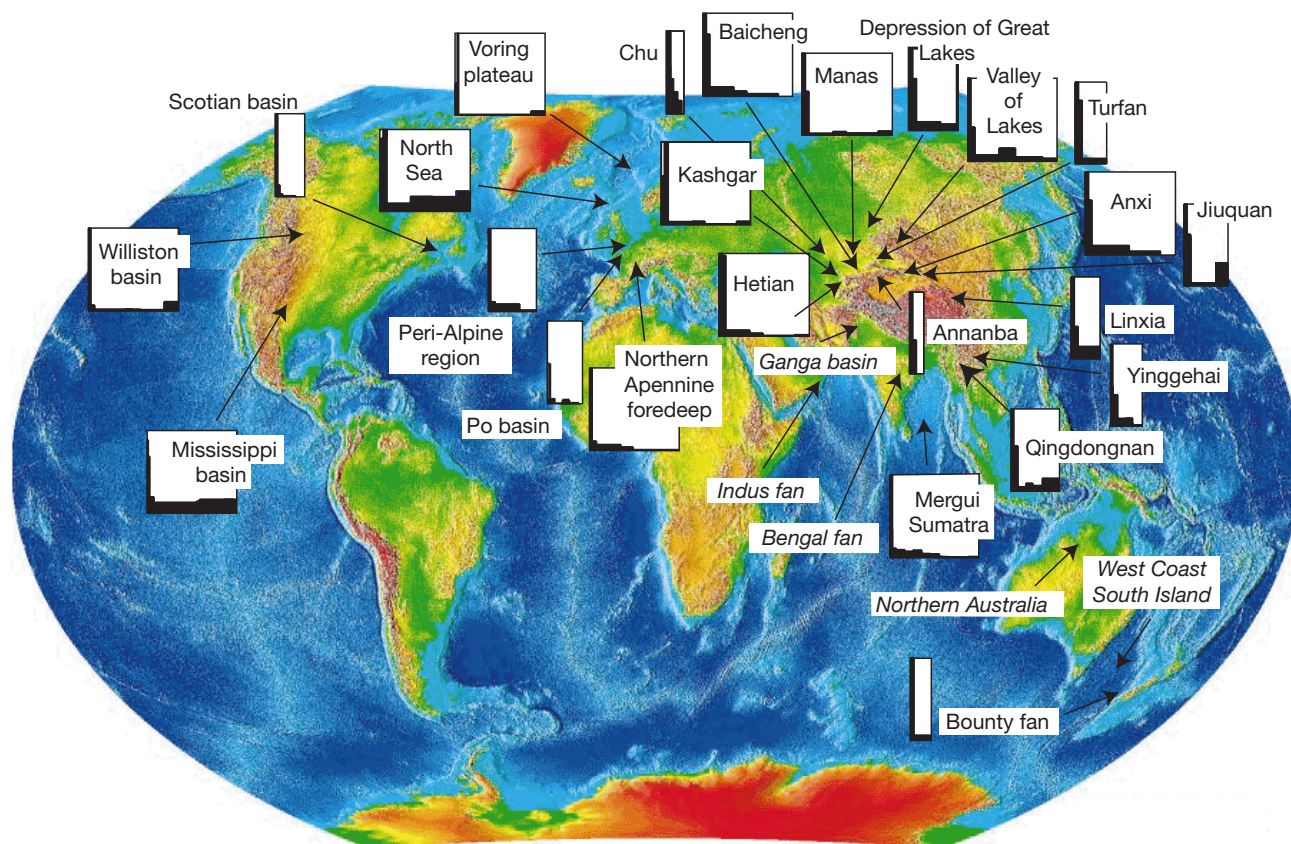
†Present address: Department of Geological Sciences, Cooperative Institute for Research in Environmental Science, Campus Box 399, University of Colorado, Boulder, Colorado 80309, USA.

increased dramatically in the past 5 Myr (Fig. 3). These authors grouped sediment into 5-Myr bins and separated it into four categories: carbonate and opaline silica (each presumably of organic origin), red clay (transported by wind) and terrigenous. The mass of terrigenous sediment deposited since 5 Myr ago is more than three times that deposited in any other 5-Myr bin. Like others, Hay *et al.*<sup>6</sup> inferred that the high rate of sediment accumulation since 5 Myr ago occurred because during periods of continental glaciation, lowered sea level allowed fluvial systems to attack continental margins<sup>6</sup>.

Consistent with this view, sediment deposited on several continental margins—such as on the Mississippi delta and its surroundings<sup>7</sup>, the North Sea<sup>8</sup>, several basins offshore southeast Asia<sup>9,10</sup>, Nova Scotia<sup>11</sup>, and the Vøring plateau off Norway<sup>12</sup>—show abrupt, several-fold increases in sediment deposition rates at 2–4 Myr ago (Figs 2, 4 and Supplementary Information). Sea-level change, however, is only one among several possible processes responsible for increases in sedimentation rate in these regions, most of which are tectonically inactive; glaciation offers another likely candidate. Rapid increases in sedimentation on the Vøring plateau (Fig. 4) and elsewhere adjacent to Norway in the Norwegian and Barents Seas logically owe their origin to glaciation, for virtually all late Cenozoic sediment seems to have accumulated after glaciation had begun<sup>12–14</sup>. Bell and Laine<sup>15</sup> estimated that pelagic sedimentation increased throughout the North Atlantic following the onset of the Laurentide glaciation. Sedimentation increased at the same time in the Williston basin in North Dakota<sup>16</sup>, far from a continental margin.

Deposition, especially of coarse material, increased abruptly in late Cenozoic times within and adjacent to much of high Asia, a

region unaffected by sea level (Figs 2, 4 and Supplementary Information). Rapid Plio-Quaternary deposition of conglomerate<sup>17–19</sup> capping a sequence of finer, older Cenozoic sediment<sup>20,21</sup> near the northern margin of Tibet has been used repeatedly to infer that the plateau rose abruptly in Quaternary times. Magnetostratigraphy<sup>18,19,22</sup> not only calibrates the Pliocene ages, once based on terrestrial fauna, assigned to underlying material, but also puts the boundary between Pliocene and Miocene as used in China near 5–6 Myr ago, close to the internationally recognized date of 5.3 Myr ago (ref. 23). The dark grey colour of the conglomerate contrasts with the reddish and orange colour of the underlying Tertiary fine-grained sediment, and implies a change from a warm, oxidizing, Tertiary environment to the colder, glacial, Quaternary period. Massive layers of poorly sorted cobbles and boulders characterize the Quaternary conglomerate near the mountains, with bedded pebble and cobble layers of conglomerate within the interiors of basins and with till in the lower parts of the conglomerate<sup>21</sup>. Spore and pollen assemblages reveal taxa, mostly *pinus*, characteristic of a cold environment. Moreover, pollen grains are much fewer than in the underlying reddish sandstone and siltstone, consistent with a colder environment with less-diverse flora than before. Finally, an abundance of heavy minerals in the conglomerate attests to less weathering of them than of the underlying sandstone and siltstone<sup>21</sup>. It appears that deposition of these conglomeratic sequences began when the climate became cold. Oxygen isotopes<sup>3</sup> from benthic foraminifers suggest that global cooling began at ~4 Myr ago (Fig. 1); magnetostratigraphy shows that loess deposition started in central and western China at ~3 Myr ago<sup>1,18,22</sup> and accelerated at that time farther east<sup>24</sup>, consistent with marked cooling and glaciation.



**Figure 2** Map of the Earth showing selected areas where sedimentation rates have increased substantially since 2–4 Myr ago. (Details are given in Fig. 4 and Supplementary Information.) For each area, a small histogram is shown. The vertical scale is normalized to

the maximum sedimentation rate in that area, and all horizontal axes are plotted at the same timescale; the longest records extend to 65 Myr ago. We show only the part of the Cenozoic for which there are measurements.



Surrounding Tibet, the average sedimentation rate since ~3–4 Myr ago has been roughly twice (or more) that of the preceding few Myr, and considerably greater than that of the preceding tens of Myr (Fig. 2 and Supplementary Information). Using magnetostratigraphy, Zheng *et al.*<sup>19</sup> recently showed that conglomeratic sedimentation began abruptly at ~3.5 Myr ago just west of Hetian (Fig. 2), with an increase in sedimentation rate of nearly ten times. This change in rate corroborates work done earlier and nearer Hetian<sup>21</sup> (Fig. 4) and at other sites along the northern margin of Tibet<sup>18,19,22,25</sup>.

Changes in sedimentation at ~3 Myr ago are not confined to the Tibetan region; similar changes, dated with magnetostratigraphy and in one case with radiometric dates on an interbedded basalt flow, can be seen in sedimentary basins that bound the Tien Shan—Kashgar, Baicheng, Turfan, Manas and Chu<sup>21,26–28</sup>—and also in Mongolia<sup>29</sup> (Figs 2 and 4). In the last, such changes are clear both in basins—the Depression of Large Lakes and the Valley of Lakes (Fig. 4)—and in the adjacent Mongolian Altay and Gobi-Altay mountain belts, respectively. Massive sequences of conglomerate overlie finer material assigned a Pliocene age, which in turn overlies thinner sequences of older Cenozoic sediment. Quaternary sedimentation rates are higher than Pliocene rates along the both margins of the Tien Shan except near Kashgar. Moreover, using fission tracks and helium isotopes to date cooling of rock in the Kyrgyz Range adjacent to the Chu basin, Bullen *et al.*<sup>30</sup> found that cooling, and presumably erosion, since ~3 Myr ago exceeded that during the previous 5–7 Myr.

A long tradition associates high rates of deposition, especially of coarse material, with recent uplift of adjacent mountain belts. Such observations, however, characterize tectonically inactive and

waning mountain belts. Hay *et al.*<sup>7</sup> suggested that the increased sedimentation rate in the Mississippi drainage basin (Fig. 4), between the Rocky Mountains and the Appalachians—where the last major deformation occurred > 50 Myr ago and > 250 Myr ago, respectively—reflected a recent, abrupt uplift of the Rockies at 2 Myr ago. The widespread belief of late Cenozoic uplift to present-day mean elevations of the Rockies derives both from glacial erosion of the previously relatively flat surface now defining the crest of the range<sup>31</sup> and from Quaternary conglomeratic sedimentation near them<sup>32</sup>. Recent applications of palaeobotany<sup>33,34</sup> to parts of the Rockies, however, indicate that the Rockies were high in mid-Cenozoic time and support the contention that climate change at ~3–4 Myr ago, rather than uplift, caused the increased erosion and sedimentation<sup>5,35</sup>.

Although perhaps not inactive, the Alps underwent their major phases of deformation in mid-Cenozoic time, if not earlier. Yet the present relief is commonly assigned a Plio-Quaternary age<sup>36</sup>. In an attempt to test whether sediment deposited in the surroundings of the Alps during Plio-Quaternary times could account for the deep valleys and present-day high relief, Guillaume and Guillaume<sup>37</sup> compiled estimates of volumes of sediment in the main surrounding basins (Fig. 4). Although these authors were forced to make some guesses of ages and of fractions of accumulated material derived from the Alps, their results suggest higher accumulation rates in Quaternary times than before (Figs 2 and 4). Apparently, half of the sediment in the Po basin is Quaternary<sup>37</sup>. Sedimentation adjacent to the Apennines also increased rapidly in late Pliocene and Quaternary times<sup>38–40</sup> (Fig. 2).

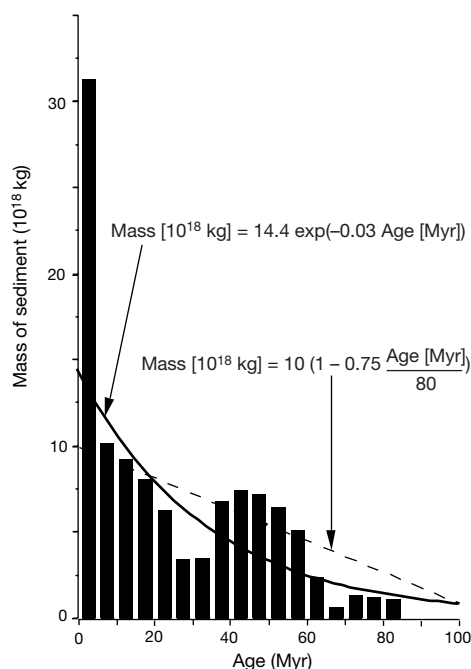
Assuming that much of the Cenozoic sediment in basins surrounding New Zealand accumulated since  $2.5 \pm 0.5$  Myr ago, and that rapid tectonic activity began at the same time, Adams<sup>41</sup> argued that present-day erosion of the Southern Alps balances the uplift of rock relative to sea level. An abrupt switch to conglomeratic sedimentation occurred northwest of the Southern Alps at that time<sup>42</sup>, and recent drilling<sup>43</sup> of the Canterbury basin and the Bounty fan off the southeast coast of New Zealand (Fig. 2) confirm rapid increases in deposition at ~2–3 Myr ago. Improved plate reconstructions<sup>44</sup>, however, require convergence and crustal shortening to have begun at 6–7 Myr ago, and therefore degrade the apparent correlation of rapid sedimentation with crustal shortening and mountain building.

### Exceptions to increases in sedimentation at ~2–4 Myr ago

Although increased sedimentation in Plio-Quaternary times has occurred throughout the world (Fig. 2), sedimentation rates did not increase at 2–4 Myr ago in all basins. For instance, although drilling on the continental shelf along the western margin of the North Atlantic suggests a Quaternary increase in sedimentation rate<sup>45</sup>, no such increase can be recognized in the adjacent, volumetrically larger, abyssal plain<sup>46</sup>.

The Ganga basin, the foredeep of the Himalaya, has recorded little change during the past 7 Myr, though grain sizes increased from relatively fine sediment to conglomerate at ~3 Myr ago<sup>47</sup>. Because this basin formed by flexure of the relatively stiff Indian plate under the load of the Himalaya thrust atop it, neither its depth nor its rate of subsidence should increase with increased sediment supplied to it. As Burbank<sup>48</sup> noted, increased erosion of the Himalaya might reduce the load flexing the Indian plate down, and hence reduce the depth of the Ganga basin and the rate of accumulation of sediment in it. Surplus sediment ought merely to result in increased transport to the Bay of Bengal, one of the world's thickest accumulations of sediment<sup>49</sup>.

Existing data do not allow a convincing demonstration of an increased sedimentation rate in the Bay of Bengal. Cochran<sup>50</sup> reported a relatively low sedimentation rate between ~8 and ~1 Myr ago at the distal edge of the Bengal fan followed by an abrupt increase at ~1 Myr ago, but this change is neither dated well



**Figure 3** Histogram of terrigenous sediment deposited in the world's oceans, compiled by Hay *et al.*<sup>6</sup>. We note the abrupt increase since ~5 Myr ago. The solid curve is an exponential fit to the data; it deviates markedly from the sedimentation rate since 5 Myr ago. The global sea floor contains nearly all the floor created in the past 5 Myr but only a fraction of that created in earlier periods. Thus, some of the greater amount of young than old sediment results from the age distribution of the sea floor. Parsons<sup>80</sup> showed that the age distribution ( $\Delta \text{area} / \Delta \text{age}$ ) of the sea floor decreases linearly with age, and the straight dashed line shows how sediment accumulating at a constant rate would be distributed, given the greater amount of young than old sea floor. The amount deposited since 5 Myr ago is more than twice that expected for a uniform rate.

nor necessarily representative of the entire basin. Métiévier *et al.*<sup>9</sup> inferred Plio-Quaternary deposition 60% more rapid than that in the late Miocene, as well as a threefold increase in accumulation rate in the Indus fan at  $\sim 2$  Myr ago. But because of Curray's<sup>49</sup> more cautious analysis of sedimentation in the Bay of Bengal, we resist using these studies<sup>9,50</sup> as additional support for a late Pliocene increase in sedimentation rates.

### The effect of climate change on erosion rates

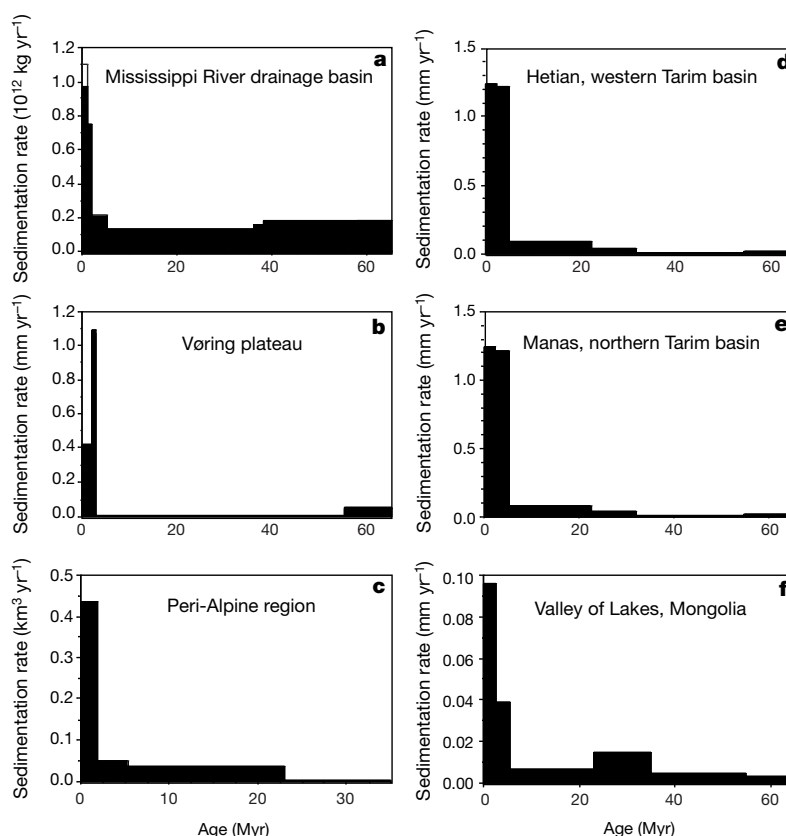
The onset of widespread glaciation at  $\sim 3$  Myr ago presents an attractive explanation for increased erosion and deposition, especially because glacial erosion rates can be higher than fluvial erosion rates<sup>51</sup>. Yet many of the areas for which geomorphologists have argued for recent uplift lie in the tropics or sub-tropics, such as southern and eastern Africa<sup>52–54</sup>, Australia, and eastern South America<sup>54</sup>. In one tropical region, northern Australia, quantitative estimates of erosion rates suggest a threefold increase in Quaternary times<sup>55</sup>. Most of Australia's landscape appears to be quite old (Mesozoic), and erosion since Cretaceous time has been very slow ( $< 1 \mu\text{m yr}^{-1}$ ; refs 56, 57). Yet Nott and Roberts<sup>55</sup> inferred an average denudation rate of  $3\text{--}4 \mu\text{m yr}^{-1}$  since approximately 0.5 Myr ago. They could not determine precisely when the erosion accelerated, but they recognized that climate change is likely to have effected the increased rate. In addition, Taylor<sup>58</sup> recognized widespread accelerated erosion in southeastern Australia in late Pliocene and

Quaternary times. Glaciers cannot have been the erosive agents in either region.

Milankovitch-scale changes in climate also have written their signature on the landscape and on the sedimentary record, without glaciation playing a direct role. The sedimentation rate in part of San Francisco Bay, California, shows large oscillations with a period of  $\sim 100$  kyr (ref. 59). As the maximum elevation of the drainage basin lies in low hills, glaciers could not have been the erosive agents. Koltermann and Gorelick<sup>59</sup> modelled these changes in sedimentation assuming that rainfall varied markedly with the 100-kyr climate cycle, as the axis of east-to-west atmospheric flow migrated north and south. Elsewhere, in tectonically active regions, the alternation between incision of canyons and widening of valleys to leave strath terraces attests to the effect of a changing climate on fluvial, not glacial, erosion<sup>60</sup>.

Apparently neither sea-level changes nor the erosive power of glaciers can account for all increases in erosion rate associated with the change from a warm equable climate (with only rare glaciers) to a cold climate (with widespread glaciers). Is there an aspect of climate change that did play a global role in accelerating erosion, or must we rely on different explanations for different regions?

One aspect of climate change apparent to all workers who study Milankovitch-scale changes is the increasing amplitude of climate variations with periods of  $\sim 20$ , 40 and 100 kyr since 3–4 Myr ago, as revealed by the marine record of  $\delta^{18}\text{O}$  (Fig. 1). Between 35 and



**Figure 4** Examples of variations in sedimentation rates, showing an abrupt increase since  $\sim 5$  Myr ago in different settings. The locations of these settings are shown in Fig. 2. Although for some examples the stratigraphic columns represent only point measurements, they characterize deposition over wide areas and demonstrate regional increases in sedimentation rates. **a**, Total mass accumulation rates for the Mississippi River basin<sup>7</sup>, between the Rocky Mountains and the Appalachians. The unshaded area for the past million years shows the estimate by Hay *et al.*<sup>7</sup> for the fraction transported into the basin by Laurentide ice sheets. **b**, Accumulation rates per unit area, without corrections for compaction, on the Vøring plateau, off Norway<sup>12</sup>. **c**, Volume accumulation rates, not corrected for compaction, in basins surrounding the Alps<sup>37</sup>. Despite much of the

sedimentation in the Black Sea being Plio-Quaternary in age and deposited in the Danube delta<sup>61</sup>, the Guillaumes<sup>37</sup> assumed that only 17% of that material was derived from the Alps. Deep-sea drilling of the Black Sea did not sample the Danube delta, but where it was drilled, sediment that accumulated since 2.5 Myr ago accounts for 50–75% of that deposited since 10 Myr ago<sup>62</sup>. **d**, Sedimentary accumulation rates, not corrected for compaction, near Hetian along the northern margin of the Tibetan plateau<sup>21</sup>. **e**, Sedimentary accumulation rates, not corrected for compaction, from near Manas along the northern margin of Tien Shan in China<sup>21</sup>. **f**, Sedimentary accumulation rates, not corrected for compaction, from the Valley of Lakes basin, which lies along the northern margin of the Gobi-Altay in Mongolia<sup>29</sup>.



30 Myr ago,  $\delta^{18}\text{O}$  in benthic organisms varied with amplitudes of 0.2–0.5‰ with periods of 100 kyr or less<sup>61</sup>, increasing to 0.5–0.6‰ between 25 and 20 Myr ago<sup>62</sup>, and then decreasing to smaller values between ~20 and 6 Myr ago<sup>63</sup>, before increasing to more than 1‰ and then nearly 2‰ since 1 Myr ago<sup>64–66</sup>. Variations with periods of 400 kyr, perhaps due to short increases (and more rarely decreases) in ice in Antarctica, contribute most of the large-amplitude variability before 20 Myr ago<sup>61,62</sup>. A similar change from a stable to an oscillating climate occurred in China, where more than 30 layers of palaeosols developed within the sequence of loess deposited since ~2.5 Myr ago indicate frequent changes between cold-dry and warm-wet environments<sup>1,67,68</sup>. Colours, grain sizes, and the composition of underlying red clay in the Chinese loess plateau, however, do not show such large climate variations from 7 to ~2.5 Myr ago<sup>67</sup>. The absence of clearly defined palaeosols in the red clay formation suggests a more stable climate between 7 and ~3 Myr ago than since ~3 Myr ago. The most important aspect of climate change for erosion may have been the change from a stable climate to one with frequent, high-amplitude variations, rather than from a warm, moist Pliocene climate to the cooler and drier ice-age climate.

Recent studies suggest that tens of thousands, and possibly millions, of years must transpire for erosional regimes to reach steady state, and therefore that landforms cannot reach equilibrium in the rapidly changing Quaternary climate<sup>69,70</sup>. Suppose climate changed from a warm-wet to a cold-dry environment. With regard to variations in climate due to precession of the Earth's axis, Gilbert<sup>71</sup> wrote that “a moist climate would tend to leach the calcareous matter from the rock, leaving an earthy soil behind, and in a succeeding drier climate the soil would be carried away.” Others<sup>72–74</sup> have developed this further by exploiting the protection afforded the Earth's surface by vegetation. Because rates of soil production decrease as soil thickness increases<sup>75</sup>, a long duration of either soil production or aridity is unlikely to lead to sustained rapid erosion, but an alternation between them takes advantage of the initial, rapid response to new, disequilibrium conditions<sup>71–74</sup>. Similarly, a climate oscillating between glacial states and periglacial states, where frost action disrupts the surface, might also accelerate erosion. Periglacial processes would fracture the surface rock and prepare it, so that glaciers could then transport broken rock. Glacial erosion seems to be dominated by quarrying, or plucking; hence preconditioning by periglacial processes should make quarrying easier than if intact rock must be extricated from the bed. Because such fracturing of rock occurs in a narrow, sub-freezing, temperature range<sup>76</sup>, only some climates will make periglacial processes effective in producing debris. Yet neither continuous periglacial conditions nor very cold conditions, such that glaciers are frozen to their beds, are likely to remove much debris.

Moreover, glaciers and rivers erode differently; the former creates wide, gently sloping valley floors, and the latter forms steep, V-shaped canyons. Using a scaling that allows incision rates from less than 1 mm yr<sup>-1</sup> to much higher rates, Harbor<sup>77</sup> concluded that the characteristic timescale for glacial erosion is ~10–100 kyr, comparable with Milankovitch periods. Rivers too must transport debris in order to erode, but much of the debris carried by a river is derived from the adjacent slopes. A river occupying a U-shaped valley, recently abandoned by a glacier, will not be burdened by debris from the neighbouring slopes, and a larger fraction of its stream power will be available to erode. The alternation between landscapes distant from glacial or fluvial equilibrium and the switch from one erosive agent to another might make erosion more rapid than if one or the other were left to do all of the erosive work.

We note that such variations between cold, dry and warmer, moister climates should be more effective than oscillations between warm, wet climates and warmer, wetter ones, because of their effects both on vegetation and on fluvial and glacial processes. Moreover, the most erosive periods in some environments might be different from those in others, and might begin at different times at different

altitudes and latitudes. Thus, if an alternation of processes leads to increased erosion and sedimentation, they need not vary perfectly synchronously throughout the Earth.

We may concoct mathematical expressions for erosion that yield enhanced rates due to an oscillating climate, with one erosive process operating in one half cycle and a second in the other. Consider first a simple linear system that approaches equilibrium exponentially, after a step change in forcing. Following Bull<sup>73</sup>, we let the erosion rate  $\dot{E}$  vary with time as

$$\dot{E} = \dot{E}_0 \exp \left[ -\frac{t}{\tau_e} \right] \sin \left[ 2\pi \frac{t}{T} \right] \quad (1)$$

for a half-cycle of oscillatory forcing, where  $\tau_e$  quantifies to time to reach an equilibrium with negligible erosion when forcing is steady, and  $T$  is the period of the oscillating amplitude of forcing. Maximum erosion per half period occurs for a vanishing ratio  $T/\tau_e$  and therefore infinitely rapid oscillations. This pattern will hold for any process whose rate decreases monotonically after a step function in forcing, including soil production<sup>75</sup>. Were erosion governed by such decaying functions, the greater amplitude of climate variability over the past 3–4 Myr should have increased erosion above that in an unchanging climate. Taken to the limit, however, it might suggest that the diurnal cycle should be the most important climate change.

Large drainage basins do not respond simultaneously to changing conditions. Bull<sup>73</sup> emphasized that landscapes responded to climate change only after a finite “reaction time”. As summarized by Bull<sup>73</sup>, R. J. Weldon demonstrated headward propagation of increased incision rate by Cajon Creek in California for several thousand years. Similarly, treatments of erosion in terms of kinematic waves propagating from one part of a watershed to another require finite intervals of time of thousands to perhaps millions of years before drainage basins can respond to changing conditions<sup>70,78</sup>. Suppose that the response to a step-function change in forcing included an increase over a finite interval, before decaying exponentially:

$$\dot{E} = \dot{E}_0 \frac{t}{\tau} \exp \left[ -\frac{t}{\tau_e} \right] \sin \left[ 2\pi \frac{t}{T} \right] \quad (2)$$

For this case, maximum erosion during a half cycle occurs for  $T \approx 3.9\tau_e$ . Because the times necessary for landscapes to respond to changes in climate are comparable to those of Milankovitch forcing, and not orders of magnitude larger or smaller, Milankovitch forcing might even show resonances with natural periods in the erosive response of a landscape to changes in forcing.

Both fluvial and glacial processes are notoriously nonlinear. Even if the thickness of the ice varied linearly with climate forcing, the rate of flow of a glacier scales as the fifth power of that thickness<sup>77</sup>, and surely erosion does not scale with the fifth root of the flow rate. Similarly, most characteristics of rivers depend in a nonlinear fashion on discharge<sup>79</sup>, which limits sediment transport. Thus, the arguments above that exploit a linear dependence of erosion on an oscillatory forcing must oversimplify the effect of changing climate on erosion. We could include a nonlinear process simply by allowing  $\tau_e$  in equation (1) to be time-dependent. (For instance, by defining  $\tau_e = \tau/[1 - (\tau/t) \ln(t/\tau)]$ , equation (1) is transformed to equation (2).) Until the physical processes that govern erosion are better understood, however, it is premature to calculate erosion rates for forcings varying differently with time.

Because late Pliocene and Quaternary times differ so much from the previous 100–200 Myr—which had an equable climate in which large changes were slow and rare—we see landforms today not only being shaped by a wide spectrum of present-day climates, but also inherited from the effects of a spectrum of different past climates. Interpreting the present-day landscape and current geomorphic processes in terms of present climate may mislead even the most careful observer. In some areas, much of the erosion that has shaped the landscape occurred during a past climate different from that of the present. The oscillation between cold-dry and warm-wet climate

may incise and denude surfaces more rapidly than would equable climates climate of any kind alone, even if erosion has occurred during only part of the past few million years. We consider that the increased sedimentation of coarser material since 2–4 Myr ago may have been caused by a climate shift. This shift was from a relatively unvarying climate, to one that oscillated between states that prepared the surface during some periods—by chemical weathering, periglacial fracturing, or other forms of mass wasting—and states that transported material.

Received 5 September 2000; accepted 2 March 2001.

- Liu, T.-S. *et al.* Loess and the Environment (China Ocean, Beijing, 1985).
- Shackleton, N. J. *et al.* Oxygen isotope calibration of the onset of ice-raftering and history of glaciation in the North Atlantic region. *Nature* **307**, 620–623 (1984).
- Maslin, M. A., Haug, G. H., Sarnthein, M. & Tiedemann, R. The progressive intensification of northern hemisphere glaciation as seen from the North Pacific. *Geol. Rdsch.* **85**, 452–465 (1996).
- Krijgsman, W., Hilgen, F. J., Raffi, I., Sierro, F. J. & Wilson, D. S. Chronology, causes and progression of the Messinian salinity crisis. *Nature* **400**, 652–655 (1999).
- Molnar, P. & England, P. Late Cenozoic uplift of mountain ranges and global climate change: chicken or egg? *Nature* **346**, 29–34 (1990).
- Hay, W. W., Sloan, J. L. & Wold, C. N. Mass/age distribution and composition of sediments on the ocean floor and the global rate of sediment subduction. *J. Geophys. Res.* **93**, 14933–14940 (1988).
- Hay, W. W., Shaw, C. A. & Wold, C. N. Mass-balanced paleogeographic reconstructions. *Geol. Rdsch.* **78**, 207–242 (1989).
- Slater, J. G. & Christie, P. A. Continental stretching: an explanation of the post-Mid-Cretaceous subsidence of the central North Sea basin. *J. Geophys. Res.* **85**, 3711–3739 (1980).
- Métivier, F., Gaudemer, Y., Tapponnier, P. & Klein, M. Mass accumulation rates in Asia during the Cenozoic. *Geophys. J. Int.* **137**, 280–318 (1999).
- Zhang, Q.-M. & Kou, C.-X. in *Sedimentary Basins of the World* Vol. 1, *Chinese Sedimentary Basins* (ed. Zhu, X.) 197–206 (Elsevier, Amsterdam 1989).
- Poag, C. W. Stratigraphic reference section for Georges Bank Basin—Depositional model for New England passive margin. *Am. Assoc. Petrol. Geol. Bull.* **66**, 1021–1041 (1982).
- Hjelstuen, B. O., Eldholm, O. & Skogseid, J. Cenozoic evolution of the northern Voring margin. *Geol. Soc. Am. Bull.* **111**, 1792–1807 (1999).
- Riis, F. Dating and measuring of erosion, uplift and subsidence in Norway and the Norwegian shelf in glacial periods. *Norsk Geol. Tidsskr.* **72**, 325–331 (1992).
- Vågenes, E., Faleide, J. I. & Gudlaugsson, S. T. Glacial erosion and tectonic uplift in the Barents Sea. *Norsk Geol. Tidsskr.* **72**, 333–338 (1992).
- Bell, M. & Laine, E. P. Erosion of the Laurentide region of North America by glacial and glacio-fluvial processes. *Quat. Res.* **23**, 154–174 (1985).
- Gerhard, L. C., Anderson, S. B., Lefever, J. A. & Carlson, C. R. Geological development, origin, and energy mineral resources of Williston Basin, North Dakota. *Am. Assoc. Petrol. Geol. Bull.* **66**, 989–1020 (1982).
- Liu, T.-S., Ding, M.-G. & Derbyshire, E. Gravel deposits on the margins of the Qinghai-Xizang Plateau, and their environmental significance. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **120**, 159–170 (1996).
- Li, J.-J., Fang, X.-M., Ma, Y.-Z. & Pan, A.-D. in *Uplift and Environmental Changes of Qinghai-Xizang Plateau in Late Cenozoic* (eds Shi, Y.-F., Li, J.-J. & Li, B.-Y.) 17–74 (Guangdong Science and Technology, Guangzhou, 1998).
- Zheng, H., Powell, C. M., An, Z., Zhou, J. & Dong, G. Pliocene uplift of the northern Tibetan Plateau. *Geology* **28**, 715–718 (2000).
- Regional Stratigraphic Tables of China—The Volume of Gansu Province* (Geological Publishing House, Beijing, 1980). (In Chinese.)
- Regional Geology of Xinjiang Uygur Autonomous Region* (Geological Memoirs Series 1, No. 32, Geological Publishing House, Beijing, 1993). (In Chinese.)
- Chen, J., Lu, Y.-C. & Ding, G.-Y. Quaternary tectonic stages in Yumen basin and western Qilian Shan. *Quat. Sci.* **1**, 263–271 (1996). (In Chinese.)
- Hilgen, F. J. & Langeris, C. G. A critical evaluation of the Miocene / Pliocene boundary as defined in the Mediterranean. *Earth Planet. Sci. Lett.* **118**, 157–179 (1993).
- Sun, D.-H., An, Z.-S., Shaw, J., Bloemendal, J. & Sun, Y.-B. Magnetostratigraphy and palaeoclimatic significance of late Tertiary aeolian sequences in the Chinese loess plateau. *Geophys. J. Int.* **134**, 207–212 (1998).
- Métivier, F., Gaudemer, Y., Tapponnier, P. & Meyer, B. Northeastward growth of the Tibet plateau deduced from balanced reconstruction of the two areas: The Qaidam and Hexi corridor basins, China. *Tectonics* **17**, 823–842 (1998).
- Deng, X.-Q., Yue, L.-P. & Teng, Z. H. A primary magnetostratigraphy study on Kuche and Xiyu formations on the edges of Tarim basin. *Acta Sedimentol. Sinica* **16**, 82–86 (1998). (In Chinese.)
- Meng, Z.-F., Li, Y.-N. & Deng, Y.-S. Paleomagnetic studies on Quaternary volcanic rock in Pulu, Xingjiang Uygur Autonomous Region. *Kexue Tongbao* **42**, 35–54 (1997). (In Chinese.)
- Bullen, M. E., Burbank, D. W., Garver, J. I. & Farley, K. A. Building the Tien Shan: Integrated thermal, structural, and topographic constraints. *J. Geol.* (submitted).
- Devyatkin, E. V. *The Cenozoic of Inner Asia* (Nauka, Moscow, 1981). (In Russian.)
- Bullen, M. E., Burbank, D. W., Abdurkhatov, K. Y. & Garver, J. I. Late Cenozoic tectonic evolution of the northwestern Tien Shan: constraints from magnetostratigraphy, detrital fission track, and basin analysis. *Geol. Soc. Am. Bull.* (in the press).
- Scott, G. R. Cenozoic history of the southern Rocky Mountains. *Mem. Geol. Soc. Am.* **144**, 227–248 (1975).
- Love, J. D. Cenozoic geology of the Granite Mountains area, central Wyoming. *Prof. Pap. US Geol. Surv.* **495-C** (1970).
- Gregory, K. M. & Chase, C. G. Tectonic significance of paleobotanically estimated climate and altitude of the late Eocene erosion surface, Colorado. *Geology* **20**, 581–585 (1992).
- Wolfe, J. A., Forest, C. E. & Molnar, P. Paleobotanical evidence of Eocene and Oligocene paleoaltitudes in midlatitude western North America. *Geol. Soc. Am. Bull.* **110**, 664–678 (1998).
- Gregory, K. M. & Chase, C. G. Tectonic and climatic significance of a late Eocene low-relief, high-level geomorphic surface, Colorado. *J. Geophys. Res.* **99**, 20141–20160 (1994).
- Triumphy, R. Paleotectonic evolution of the central and western Alps. *Geol. Soc. Am. Bull.* **71**, 843–908 (1960).
- Guillaume, A. & Guillaume, S. L'érosion dans les Alpes au Plio-Quaternaire et au Miocène. *Ecol. Geol. Helv.* **75**, 247–268 (1982).
- Ori, G. G., Roveri, M. & Vannoni, F. in *Foreland Basins* (eds Allen, P. A. & Homewood, P.) 183–198 (Spec. Publ. No. 8, International Association of Sedimentologists, Blackwell, Oxford, 1986).
- Pieri, M. & Mattavelli, L. Geologic framework of Italian Petroleum resources. *Am. Assoc. Petrol. Geol. Bull.* **70**, 103–130 (1986).
- Ricci Lucchi, F. in *Foreland Basins* (eds Allen, P. A. & Homewood, P.) 105–139 (Spec. Publ. No. 8, International Association of Sedimentologists, Blackwell, Oxford, 1986).
- Adams, J. Contemporary uplift and erosion of the Southern Alps, New Zealand. *Geol. Soc. Am. Bull.* **91**, 1–114 (1980).
- Nathan, S. *et al.* *Cretaceous and Cenozoic Sedimentary Basins of the West Coast Region, South Island, New Zealand* (New Zealand Geological Survey Basin Studies 1, Department of Scientific and Industrial Research, Wellington, 1986).
- Shipboard Scientific Party. Leg 181 summary: Southwest Pacific paleoceanography. *Proc. ODP Init. Rep.* **181**, 1–80 (1999).
- Walcott, R. I. Modes of oblique compression: Late Cenozoic tectonics of the South Island of New Zealand. *Rev. Geophys.* **36**, 1–26 (1998).
- Poag, C. W. in *Geologic Evolution of the United States Atlantic Margin* (ed. Poag, C. W.) 217–264 (Van Nostrand Reinhold, New York, 1985).
- Mountain, G. S. & Tucholke, B. E. in *Geologic Evolution of the United States Atlantic Margin* (ed. Poag, C. W.) 293–341 (Van Nostrand Reinhold, New York, 1985).
- Burbank, D. W., Beck, R. A. & Mulder, T. in *The Tectonic Evolution of Asia* (eds Yin, A. & Harrison, T. M.) 149–188 (Cambridge Univ. Press, Cambridge, 1996).
- Burbank, D. W. Causes of recent Himalayan uplift deduced from deposited patterns in the Ganges basin. *Nature* **357**, 680–683 (1992).
- Curry, J. R. Sediment volume and mass beneath the Bay of Bengal. *Earth Planet. Sci. Lett.* **125**, 371–383 (1994).
- Cochran, J. R. Himalayan uplift, sea level, and the record of Bengal Fan sedimentation at the ODP Leg 116 sites. *Proc. ODP Sci. Res.* **116**, 397–414 (1990).
- Hallet, B., Hunter, L. & Bogen, J. Rates of erosion and sediment evacuation by glaciers: A review of field data and their implications. *Glob. Planet. Change* **12**, 213–235 (1996).
- Baker, B. H., Mohr, P. A. & Williams, L. A. J. Geology of the eastern rift system. *Spec. Pap. Geol. Soc. Am.* **136** (1972).
- Saggerson, E. P. & Baker, B. H. Post-Jurassic erosion-surfaces in eastern Kenya and their deformation in relation to rift structure. *Q. J. Geol. Soc. Lond.* **121**, 51–72 (1965).
- King, L. C. *The Morphology of the Earth* (Hafner, New York, 1967).
- Nott, J. & Roberts, R. G. Time and process rates over the past 100 m.y.: A case for dramatically increased landscape denudation rates during the late Quaternary in northern Australia. *Geology* **24**, 883–887 (1996).
- Twidale, C. R. On the survival of paleolandforms. *Am. J. Sci.* **276**, 77–95 (1976).
- Nott, J. The antiquity of landscapes on the north Australian craton and the implications for theories of long-term landscape evolution. *J. Geol.* **103**, 19–32 (1995).
- Taylor, G. A brief Cainozoic history of the Upper Darling Basin. *Proc. R. Soc. Vict.* **90**, 53–59 (1978).
- Koltermann, C. E. & Gorelick, S. M. Paleoclimatic signature in terrestrial flood deposits. *Science* **256**, 1775–1782 (1992).
- Molnar, P. *et al.* Quaternary climate change and the formation of river terraces across growing anticlines on the north flank of the Tien Shan, China. *J. Geol.* **102**, 583–602 (1994).
- Zachos, J. C., Quinn, T. M. & Salamy, K. A. High-resolution (10<sup>4</sup> years) deep-sea foraminiferal stable isotope records of the Eocene-Oligocene climate transition. *Paleoceanography* **11**, 251–266 (1996).
- Zachos, J. C., Flower, B. P. & Paul, H. A. Orbital paced climate oscillations across the Oligocene/Miocene boundary. *Nature* **388**, 567–570 (1997).
- Shackleton, N. J. & Hall, M. A. The late Miocene stable isotope record, Site 926. *Proc. ODP Sci. Res.* **154**, 367–373 (1997).
- Bickert, T., Curry, W. B. & Wefer, G. Late Pliocene to Holocene (2.6–0 Ma) western equatorial Atlantic deep-water circulation: Inferences from benthic stable isotopes. *Proc. ODP Sci. Res.* **154**, 239–253 (1997).
- Billups, K., Ravelo, A. C. & Zachos, J. C. Early Pliocene deep-water circulation: Stable isotope evidence for enhanced northern component deep water. *Proc. ODP Sci. Res.* **154**, 319–330 (1997).
- Tiedemann, R. & Franz, S. O. Deep-water circulation, chemistry, and terrigenous sediment supply in the equatorial Atlantic during the Pliocene, 3.3–2.6 Ma and 5–4.5 Ma. *Proc. ODP Sci. Res.* **154**, 299–318 (1997).
- Ding, Z.-L., Sun, J.-M. & Yang, S.-L. Magnetostratigraphy and grain size record of a thick red clay-loess sequence at Lingtai, the Chinese loess plateau. *Quat. Sci.* **1**, 86–94 (1998). (In Chinese.)
- Kukla, G. *et al.* Pleistocene climates in China dated by magnetic susceptibility. *Geology* **16**, 811–814 (1988).
- Fernandes, N. F. & Dietrich, W. E. Hillslope evolution by diffusive processes: The timescale for equilibrium adjustments. *Wat. Resour. Res.* **33**, 1307–1318 (1997).
- Whipple, K. X. Fluvial landscape response time: How plausible is steady-state denudation. *Am. J. Sci.* (in the press).
- Gilbert, G. K. Rhythms and geologic time. *Proc. Am. Assoc. Adv. Sci.* **49**, 1–19 (1900).
- Knox, J. C. Valley alluviation is southeastern Wisconsin. *Annu. Assoc. Am. Geogr.* **62**, 401–410 (1972).
- Bull, W. B. *Geomorphic Responses to Climatic Change* (Oxford Univ. Press, Oxford, 1991).
- Tucker, G. E. & Slingerland, R. Drainage responses to climate change. *Wat. Resour. Res.* **33**, 2031–2047 (1997).
- Heimsath, A. M., Dietrich, W. E., Nishiizumi, K. & Finkel, R. C. The soil production function and landscape evolution. *Nature* **388**, 358–361 (1997).
- Walder, J. S. & Hallet, B. A theoretical model of the fracture of rock during freezing. *Geol. Soc. Am. Bull.* **96**, 336–346 (1985).
- Harbor, J. M. Numerical modeling of the development of U-shaped valleys by glacial erosion. *Geol. Soc. Am. Bull.* **104**, 1364–1375 (1992).
- Whipple, K. X. & Tucker, G. E. Dynamics of stream-power river incision model: Implications for



height limit of mountain ranges, landscape response timescales, and research needs. *J. Geophys. Res.* **104**, 17661–17674 (1999).

79. Leopold, L. B., Wolman, M. G. & Miller, J. P. *Fluvial Processes in Geomorphology* (Freeman, San Francisco, 1964).
80. Parsons, B. Causes and consequences of the relation between area and age of the ocean floor. *J. Geophys. Res.* **87**, 437–448 (1982).
81. Letouzey, J., Gonnard, R., Montadert, L., Kristchev, K. & Dorkel, A. Black Sea: Geological setting and recent deposits distribution from seismic reflection data. *Init. Rep. DSDP* **42**(2), 1077–1084 (1978).
82. Hsü, K. J. Correlation of Black Sea sequences. *Init. Rep. DSDP* **42**(2), 489–497 (1978).

**Supplementary information** is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

## Acknowledgements

This work was stimulated by a lecture by T.-L. Liu in 1992; in 1998, S. Thompson pointed out that a steady-state geomorphic regime would not be established in an environment undergoing large, rapid climate changes. We thank T. Bickert and W. B. Curry for the unpublished data shown in Fig. 1, W. B. Curry for assembling these data for us, Z.-L. Ding for guidance with the Quaternary climate history of China, C. Duncan for help with figures, and J. R. Curray and B. Hallet for constructive criticisms. P.-Z. Zhang is supported by the National Science Foundation of China.

Correspondence and requests for materials should be addressed to P.M. (e-mail: [molnar@terra.colorado.edu](mailto:molnar@terra.colorado.edu)).

# Resolution of distinct rotational substeps by submillisecond kinetic analysis of F<sub>1</sub>-ATPase

Ryohei Yasuda<sup>\*†||</sup>, Hiroyuki Noji<sup>\*</sup>, Masasuke Yoshida<sup>‡</sup>, Kazuhiko Kinoshita Jr<sup>†\*</sup> & Hiroyasu Itoh<sup>§\*</sup>

<sup>\*</sup>CREST 'Genetic Programming' Team 13, Teikyo University Biotechnology Center 3F, Nogawa 907, Miyamae-Ku, Kawasaki 216-0001, Japan

<sup>†</sup>Department of Physics, Faculty of Science and Technology, Keio University, Yokohama 223-8522, Japan

<sup>‡</sup>Chemical Resources Laboratory, Tokyo Institute of Technology, Yokohama 226-8503, Japan

<sup>§</sup>Tsukuba Research Laboratory, Hamamatsu Photonics KK, Tokodai, Tsukuba 300-2635, Japan

The enzyme F<sub>1</sub>-ATPase has been shown to be a rotary motor in which the central  $\gamma$ -subunit rotates inside the cylinder made of  $\alpha_3\beta_3$  subunits. At low ATP concentrations, the motor rotates in discrete 120° steps, consistent with sequential ATP hydrolysis on the three  $\beta$ -subunits. The mechanism of stepping is unknown. Here we show by high-speed imaging that the 120° step consists of roughly 90° and 30° substeps, each taking only a fraction of a millisecond. ATP binding drives the 90° substep, and the 30° substep is probably driven by release of a hydrolysis product. The two substeps are separated by two reactions of about 1 ms, which together occupy most of the ATP hydrolysis cycle. This scheme probably applies to rotation at full speed (~130 revolutions per second at saturating ATP) down to occasional stepping at nanomolar ATP concentrations, and supports the binding-change model for ATP synthesis by reverse rotation of F<sub>1</sub>-ATPase.

The ATP synthase is an enzyme ubiquitous in bacteria, plants and animals, which synthesizes ATP from ADP and inorganic phosphate using proton flow through a membrane<sup>1–3</sup>. F<sub>1</sub>, a water-soluble portion of the ATP synthase, is the site of ATP synthesis, whereas protons flow through the membrane-embedded F<sub>0</sub> portion. At least *in vitro*, F<sub>1</sub> can hydrolyse ATP to pump protons through the F<sub>0</sub> portion in the reverse direction. Isolated F<sub>1</sub> only hydrolyses ATP, and is called F<sub>1</sub>-ATPase. Its subunit composition is  $\alpha_3\beta_3\gamma\delta\epsilon$ .

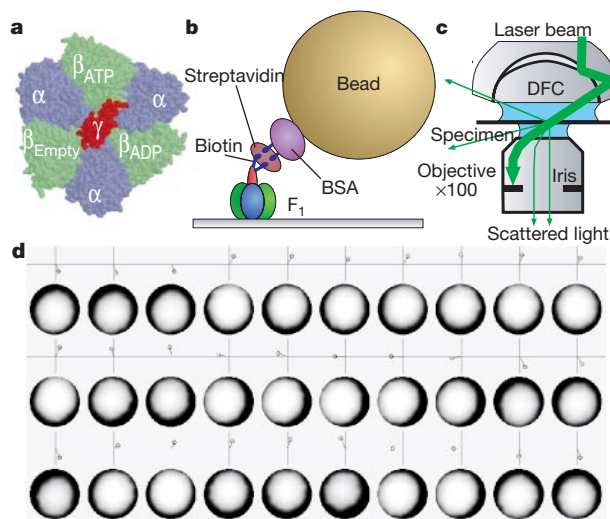
The prevailing view is that ATP hydrolysis/synthesis in F<sub>1</sub> is coupled to proton flow in F<sub>0</sub> through the rotation of a common shaft, of which the  $\gamma$ -subunit of F<sub>1</sub> is a part. This rotational coupling mechanism was initially proposed by Boyer<sup>1–3</sup>, and by others<sup>4–6</sup>. Later, a crystal structure of F<sub>1</sub> showed that a rod-shaped  $\gamma$ -subunit is surrounded by a cylinder made of three  $\alpha$ - and three  $\beta$ -subunits, arranged alternately<sup>7</sup> (Fig. 1a). An analogue of ATP, ADP and none were bound to the three  $\beta$ -subunits, indicating that sequential ATP hydrolysis on the three  $\beta$ -subunits would indeed induce rotation of the central, asymmetrical  $\gamma$ -subunit. Rotation of the  $\gamma$ -subunit in an isolated F<sub>1</sub> during ATP hydrolysis has been demonstrated experimentally by various methods<sup>8–10</sup>.

We have visualized the rotation of the  $\gamma$ -subunit under an optical microscope by fixing F<sub>1</sub> on a surface and attaching an actin filament to the  $\gamma$ -subunit as a marker of its orientation<sup>10</sup>. At nanomolar ATP, the actin filament rotated in discrete 120° steps<sup>11</sup>, consistent with the pseudo-three-fold symmetrical structure<sup>7</sup> of F<sub>1</sub>. The rotation rate was close to one-third of the rate of ATP hydrolysis in solution, suggesting that one ATP molecule is consumed per 120° step<sup>11</sup>. At high ATP concentrations, however, the actin rotation was smooth rather than stepwise, and the rotation was much slower than ATP hydrolysis. Viscous friction imposed on the actin filament prevented fast rotation of F<sub>1</sub> and obscured the stepping behaviour. Here we have used a smaller marker, a colloidal gold bead of 40-nm diameter, for which the viscous friction is 10<sup>–3</sup> to 10<sup>–4</sup> times that for actin (Fig. 1b). In the resultant high-speed rotation, we were able to resolve substeps. In this study we investigate the magnitudes,

speeds, and timings of the substeps, and we also look at: (1) the motor speeds at no load; (2) whether the motor uses different rotary mechanisms at low and high speeds; (3) which parts of hydrolysis reactions drive the substeps; and (4) what structural changes may underlie the substeps.

## Full-speed rotation with 40-nm beads

Bead rotation was imaged by laser dark-field microscopy<sup>12</sup> (Fig. 1c),



**Figure 1** Observation of F<sub>1</sub> rotation. **a**, Atomic structure<sup>7</sup> of F<sub>1</sub>-ATPase viewed from the F<sub>0</sub> side (top in **b**). **b**, Side view of the observation system. The 40-nm bead gave a large enough optical signal that warranted a submillisecond resolution; but the bead was small enough not to impede the rotation. **c**, Laser dark-field microscopy for observation of gold beads. Only light scattered by the beads exited the objective and was detected. DFC, dark-field condenser. **d**, Sequential images of a rotating bead at 2 mM ATP. Images are trimmed in circles (diameter 370 nm) to aid identification of the bead position; centroid positions are shown above the images at  $\times 3$  magnification. The interval between images is 0.5 ms.

<sup>||</sup>Present address: Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA.



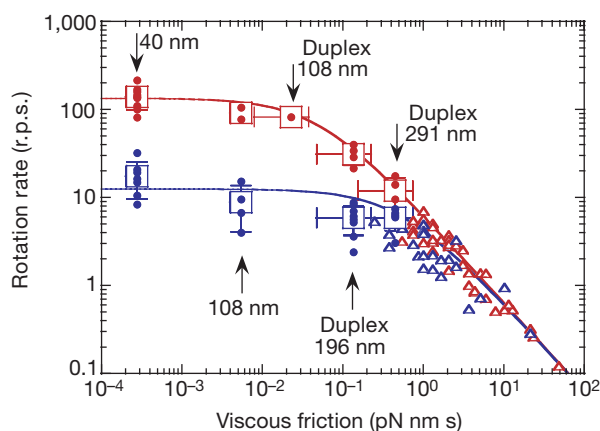
and recorded on a fast-framing charge-coupled-device (CCD) camera at speeds up to 8,000 frames per s. The 40-nm bead appeared as a spot of diffraction-limited size ( $\sim 300$  nm; Fig. 1d). When a bead is attached obliquely (Fig. 1b), rotation of the  $\gamma$ -subunit will result in a circular movement of the bead image. Some beads showed rotation (Fig. 1d; movies in Supplementary Information), and motions of these beads were analysed by calculating the centroid of the bead image<sup>13</sup>. The rotation diameter of bead centroid ranged between 25–55 nm. Diameters up to  $\sim 60$  nm are possible for the height of  $F_1$  of  $\sim 10$  nm and the linker lengths of  $\sim 5$  nm for streptavidin<sup>14</sup> and  $\sim 10$  nm for BSA<sup>15</sup> (Fig. 1b). Rotation was stepwise at all ATP concentrations examined (see below).

To see whether the friction on the 40-nm bead impeded  $F_1$  rotation, we varied the frictional load by attaching single or duplex polystyrene beads (108, 196 or 291 nm) to the  $\gamma$ -subunit. At both 2 mM and 2  $\mu$ M ATP (Fig. 2; red and blue circles, respectively), time-averaged rotation rates showed saturation behaviour at small friction. Maximal rotation rate depended on ATP concentration [ATP], but increasing [ATP] beyond 2 mM did not accelerate rotation (see below). Thus, at 2 mM ATP, the 40-nm bead rotated at the full speed of the  $F_1$  motor, which was 134 revolutions per second (r.p.s.) at 23°C; the bead was not an impeding load for  $F_1$ . At saturating speeds, all beads rotated stepwise. On the load-dependent portions in Fig. 2, however, bead rotation was smooth, as was rotation of actin at these ATP concentrations<sup>11</sup>. The load dependence of actin rotation<sup>11</sup> (Fig. 2; triangles) is consistent with the bead assay.

### One rotary mechanism at all speeds

The time-averaged rate of rotation showed simple Michaelis–Menten dependence on [ATP] (Fig. 3; the maximal rate at infinite [ATP],  $V_{\max} = 129$  r.p.s.; Michaelis constant,  $K_m = 15$   $\mu$ M), suggesting that one mechanism accounts for rotation in the nM–mM range. This idea is corroborated by observations<sup>11,16,17</sup> that the torque and its angle dependence, as well as mechanical work done in a 120° step, are independent of [ATP] over the nM–mM range. Also, the apparent rate of ATP binding,  $k_{\text{on}}^{\text{ATP}}$ , given by  $3V_{\max}/K_m$  of  $(2.6 \pm 0.5) \times 10^7 \text{ M}^{-1} \text{ s}^{-1}$  agrees with previous estimates based on the analysis of step intervals at nanomolar ATP<sup>11,18</sup>.

As seen in Fig. 3, the rotation rate was close to one-third of

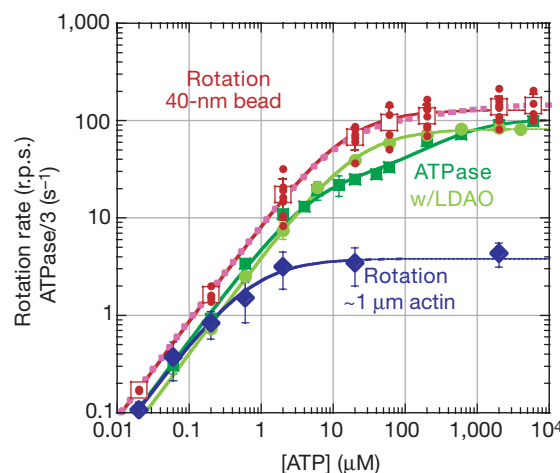


**Figure 2** Relationship between rate of bead rotation and viscous friction on the bead. Circles, the average rate for a bead calculated over at least 20 consecutive revolutions; squares, the average over different beads (vertical error bars indicate s.d.). ATP at 2 mM and 2  $\mu$ M is indicated by red and blue colours, respectively. The abscissa is the rotational frictional drag coefficient  $\xi$  calculated as in Methods. Possible range of  $\xi$  for each bead is shown by the size of the squares or associated horizontal error bars. For comparison, rotation rates for an actin filament attached to the  $\gamma$ -subunit<sup>11</sup> are also plotted (triangles). Lines show fits with the rate expected for a motor producing a constant torque<sup>11</sup>:  $(1/V_{\text{load}} + 2\pi\xi/N)^{-1}$  where  $N = 40$  pN nm (assumed torque) and  $V_{\text{load}} = 12.5 \pm 1.0$  r.p.s. for 2  $\mu$ M ATP and  $134 \pm 3$  r.p.s. for 2 mM ATP (s.e.m.).

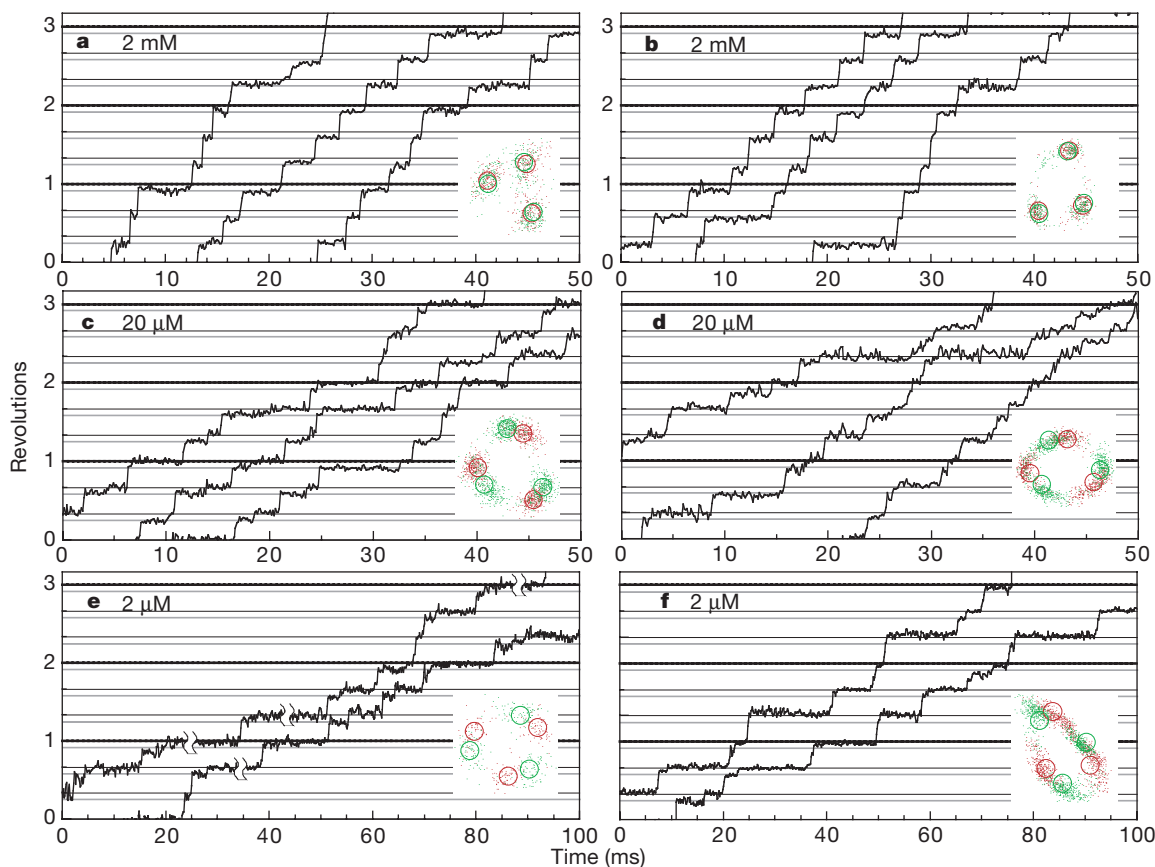
the rate of ATP hydrolysis for bead-free  $F_1$  in solution, supporting the contention that one ATP molecule is consumed per 120° rotation<sup>11,16–18</sup>. The hydrolysis rate, however, was lower, particularly around 50  $\mu$ M. A probable cause is MgADP inhibition:  $F_1$  is stochastically inactivated during ATP hydrolysis, when it binds MgADP tightly<sup>17,19,20</sup>. Although we started with nucleotide-free  $F_1$ , some inactivation may have proceeded during the mixing time of  $\sim 2$  s. Indeed, the rate of inactivation increases with [ATP] and reaches  $\sim 0.3 \text{ s}^{-1}$  at  $> 10$   $\mu$ M ATP<sup>20</sup>, the position of the concavity in Fig. 3. Higher activity at still higher [ATP] is accounted for by binding of ATP to non-catalytic  $\alpha$ -subunits, which tends to restore the hydrolysis activity<sup>19</sup>. Lauryldodecylamine oxide (LDAO), a suppressor of the MgADP inhibition<sup>19</sup>, produced hydrolysis kinetics parallel to the rotation kinetics, although  $V_{\max}/3$  ( $82 \text{ s}^{-1}$ ) was only  $\sim 60\%$  of  $V_{\max}$  for rotation (Fig. 3).

### The 120° step consists of 90° and 30° substeps

At 8,000 frames per s, steps were clearly resolved in the rotation of 40-nm beads, even at saturating ATP. At 2 mM ATP, only 120° steps were seen (Fig. 4a, b), whereas at 20  $\mu$ M or 2  $\mu$ M ATP, the 120° step was further split into roughly 90° and 30° substeps (Fig. 4c–f, where each panel shows a continuous record). We call the interval between a 30° substep and a 90° substep a ‘0° dwell’ and the interval between 90° and 30° substeps a ‘90° dwell’. In Fig. 4c–f, 0° dwells fall on black horizontal lines that are separated from each other by 120°, and 90° dwells fall on grey lines that are 30° below the black lines. The 90° dwells were about a few ms in duration, on average, both at 2 and 20  $\mu$ M ATP, whereas 0° dwells became longer at 2  $\mu$ M. The implication is that  $F_1$  waits for the arrival of ATP during the 0° dwell, which is terminated by a 90° substep induced by ATP binding. The subsequent 90° dwell is for a process or processes independent of [ATP]. This scheme predicts that, at [ATP]  $\sim K_m = 15$   $\mu$ M, 0° and 90° dwells have approximately equal lengths, as was observed



**Figure 3** Comparison of rotation and hydrolysis rates. Red circles, time-averaged rotation rate for individual 40-nm beads. Red squares, rotation rate averaged over different beads. Dark green squares, one-third of the initial rate of ATP hydrolysis. Light green circles, one-third of the rate of ATP hydrolysis in the presence of LDAO. Blue diamonds, rotation rate for an actin filament attached to the  $\gamma$ -subunit<sup>11</sup>. Standard deviations greater than the symbol size are shown in bars ( $n \geq 2$ ). Curves show fits with Michaelis–Menten kinetics,  $V = V_{\max}[\text{ATP}]/(K_m + [\text{ATP}])$ , where  $V_{\max}$  and  $K_m$  are  $129 \pm 9$  r.p.s. and  $15 \pm 2$   $\mu$ M for bead rotation (red),  $4.0 \pm 0.3$  r.p.s. and  $0.7 \pm 0.1$   $\mu$ M for actin rotation (blue), and  $247 \pm 9 \text{ s}^{-1}$  and  $19 \pm 1$   $\mu$ M for hydrolysis in the presence of LDAO (light green). Fits with two  $K_m$  values,  $V = (V_{\max1}K_{m2}[\text{ATP}] + V_{\max2}[\text{ATP}]^2)/([K_{m1}[\text{ATP}] + K_{m1}K_{m2}] + K_{m1}K_{m2})$ , are also shown, where  $V_{\max1} = 85 \pm 9 \text{ s}^{-1}$ ,  $K_{m1} = 5.2 \pm 0.7$   $\mu$ M,  $V_{\max2} = 306 \pm 22 \text{ s}^{-1}$ , and  $K_{m2} = 393 \pm 147$   $\mu$ M for hydrolysis without LDAO (dark green), and  $V_{\max1} = 109 \pm 30$  r.p.s.,  $K_{m1} = 12 \pm 4$   $\mu$ M,  $V_{\max2} = 149 \pm 32$  r.p.s., and  $K_{m2} = 682 \pm 2768$   $\mu$ M for bead rotation (dashed pink). The latter does not show improvement over the simple fit in red. Values are means  $\pm$  s.e.m.



**Figure 4** Unfiltered time courses of stepping rotation of 40-nm beads at varying [ATP]. **a, b**, 2 mM; **c, d**, 20 μM; **e, f**, 2 μM [ATP]. All curves in a panel are continuous; later curves are shifted, to save space. Grey horizontal lines are placed 30° below black lines. In **e**, some of the long dwells are cut short. Insets, positions of a bead within 0.25–0.5 ms

at 20 μM ATP. At 2 mM ATP, the expected rate of ATP binding is  $(2.6 \times 10^7 \text{ M}^{-1} \text{ s}^{-1} \times 2 \text{ mM}) \approx 5 \times 10^4 \text{ s}^{-1}$ . Then, 0° dwells will be ~0.02 ms and will not be detected at the current resolution. The absence of substeps in Fig. 4a, b is thus explained. We place dwells in Fig. 4a, b on grey lines, because they must be 90° dwells according to this explanation (see also Fig. 7).

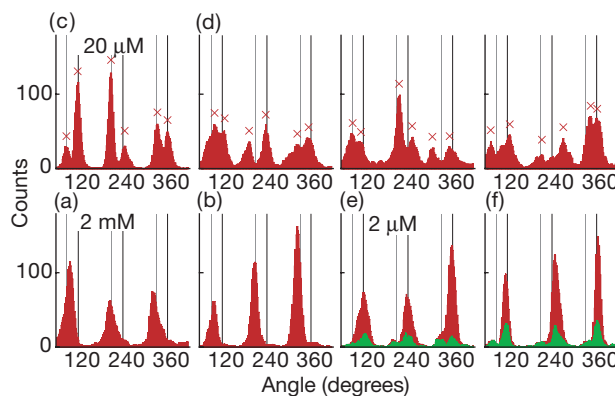
The 30° substeps were not always clear, but we could easily locate, in rotation records, steps that spanned most of a 120° interval (~90° or ~120° step). Positions of the bead centroid in 0.25–0.5-ms intervals before (red) and after (green) these main steps are shown in the insets of Fig. 4. At 2 μM and 20 μM ATP, red and green spots are separated by ~30°, showing the presence of substeps, whereas spots overlap with each other at 2 mM ATP. The traces are distorted, presumably because of oblique rotation on an obliquely situated  $F_1$ . Circles on each trace are the projection of three equally spaced pairs of dwell positions on a circular trajectory oblique to the surface. A search for the best fit with the observed traces show the separation between red and green circles to be  $29^\circ \pm 7^\circ$  (mean  $\pm$  s.d.) for 13 runs at 2, 6 and 20 μM ATP; and  $4^\circ \pm 3^\circ$  for 7 runs at 2 and 6 μM.

Figure 5 shows histograms of angular positions. Separations of peaks at 20 μM ATP (crosses) averaged  $35^\circ \pm 13^\circ$  (mean  $\pm$  s.d. for 15 peak pairs). Substeps are less obvious at 2 μM, but histograms for the 2-ms intervals before and after the main steps (green) show similar peak separations. Taking these and other experimental uncertainties into account, we estimate the substep sizes to be within  $90^\circ \pm 10^\circ$  and  $30^\circ \pm 10^\circ$ .

## Steps are fast

Consecutive steps in a rotation record at 2 mM ATP are super-

before (red) and after (green) the main (90° or 120°) steps; runs lasting 0.5 s (2 mM) or 2 s (2 μM and 20 μM) were analysed. Circles indicate projection of ~0° and ~90° dwell points on an obliquely situated circular trajectory that best fit the data. Angles in the time courses and in Fig. 5 are those on the oblique circle.



**Figure 5** Histograms of angular positions over 0.5 s runs. Labels (a–f) are from records of which Fig. 4, **a–f** are a part. Each time course was passed through a five-point median filter, and its histogram was calculated with 2° bins. The histogram was then averaged over 10° intervals. Green parts (**e** and **f**) indicate 2 ms before and after main steps. Crosses indicate peaks identified by eye. To assess the baseline noise in raw data, we also constructed unfiltered, unaveraged histograms at 2 μM ATP with 2° bins (not shown). The histograms gave three peaks, of which the half width at 1/e height was estimated by fitting each peak with a gaussian curve; the half widths averaged  $18^\circ \pm 7^\circ$  (mean  $\pm$  s.d. for 15 peaks).



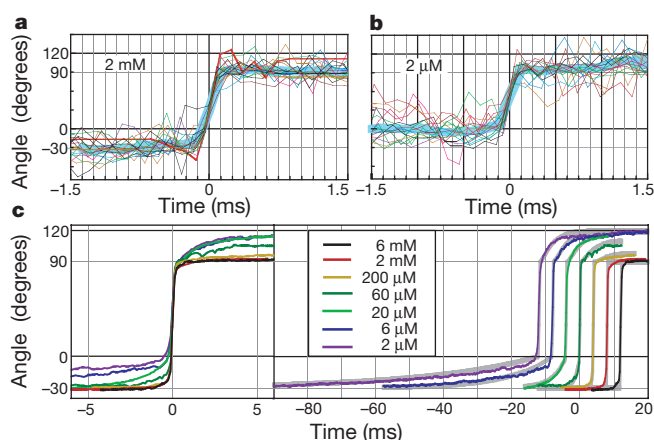
imposed in Fig. 6a. The average (thick cyan line) shows that a whole 120° step completes within 0.25 ms (two frames) at saturating ATP. This value is an upper boundary because faster transients are unresolved with the camera that we used. Thus, time for the mechanical stepping (the times needed to reorientate the  $\gamma$ -subunit through 90° and 30°) occupies < 10% of the ATPase cycle time. The 90° substeps at 2  $\mu$ M ATP are also within 0.25 ms (Fig. 6b). Stepping is fast, the instantaneous speed being well above 1,000 r.p.s., whether [ATP] is above or below  $K_m$  for rotation.

### Substeps of 90° by ATP binding and 30° by product release

Figure 6c shows averages of all steps observed at indicated [ATP]. Presence of distinct and fast  $\sim 90^\circ$  substeps is clear at all [ATP] < 60  $\mu$ M, although whether the remaining  $\sim 30^\circ$  is also stepwise cannot be judged in this figure. Fit with grey lines indicates the substep size to be  $90.2^\circ \pm 0.3^\circ$ .

The averaging was made after steps other than the central one were eliminated from each step record, such that it started with a near  $-30^\circ$  dwell and ended with a  $\sim 120^\circ$  dwell. When the last dwell at  $120^\circ$  was too short to be distinguished, the previous dwell at  $\sim 90^\circ$  was extended to the right edge of the figure (see Methods). Thus, the portion of the curves between  $-30^\circ$  and  $0^\circ$  reflects the distribution of dwell times at  $0^\circ$  that were started at the end of a substep from  $-30^\circ$  to  $0^\circ$ , and terminated by a central  $90^\circ$  substep. The dwell is [ATP]-dependent, and can be explained by termination by ATP binding at the rate  $k_{on}^{ATP} = 3.0 \times 10^7 \text{ M}^{-1} \text{ s}^{-1}$  estimated from Fig. 8 (Fig. 6c; grey lines). The kinetics above  $90^\circ$ , in contrast, is [ATP]-independent, except that the amplitude decreases with [ATP]. The decrease is accounted for by the fact that, at [ATP] >  $K_m$ , the dwell at  $\sim 120^\circ$  becomes too short and is not represented in our averaging procedure. As shown (grey lines), the kinetics are in accord with the scheme in Fig. 7a where two  $\sim 1$ -ms reactions govern the dwell at  $\sim 90^\circ$  (the two reactions are deduced from Fig. 8).

Our proposed scheme is summarized in Fig. 7a. ATP binding drives a  $90^\circ$  substep (A  $\rightarrow$  B). Next are hydrolysis reactions that are mechanically silent (B, C). Eventually, the last hydrolysis product of the previously bound ATP (ADP, phosphate, or both) is released, accompanying a  $30^\circ$  substep (C  $\rightarrow$  A') and resetting the system to the initial A state except for the  $120^\circ$  rotation that has taken place. At



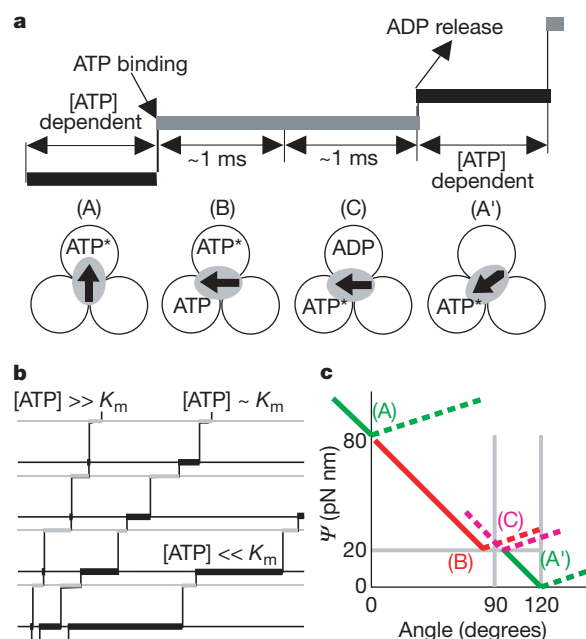
**Figure 6** Kinetics of substeps. **a, b**, Eighteen consecutive steps and their average (thick cyan line) in a rotation record at 2 mM (**a**) or 2  $\mu$ M (**b**) ATP (see Methods for the averaging procedure). **c**, Steps in several runs at indicated [ATP], averaged as in **a** and **b**. The averaging procedure retained all dwells at  $\sim 0^\circ$  position in the ordinate, but some dwells at  $\sim 120^\circ$  were converted to a horizontal line at  $\sim 90^\circ$  position when the substep from  $90^\circ$  to  $120^\circ$  was contiguous (within 0.25 ms) with the next substep from  $120^\circ$  to  $210^\circ$ . This is why the curves at high [ATP] do not rise much beyond the  $90^\circ$  line. The curves are reproduced on the right; superimposed grey lines represent fits with theoretical curves based on the scheme in Fig. 7a (see Methods). The best fit was obtained with the size of the  $30^\circ$  substep,  $A_{30^\circ}$ , of  $29.8^\circ \pm 0.3^\circ$  (mean  $\pm$  s.e.m.).

high [ATP], another  $90^\circ$  substep occurs immediately, and thus the two substeps are not resolved (Fig. 7b)—this is why we propose that both of the hydrolysis products must have been released by the end of a  $30^\circ$  substep. At [ATP]  $\sim K_m$ , ATP binding and hydrolysis take similar time, and  $90^\circ$  and  $30^\circ$  substeps are equally spaced.

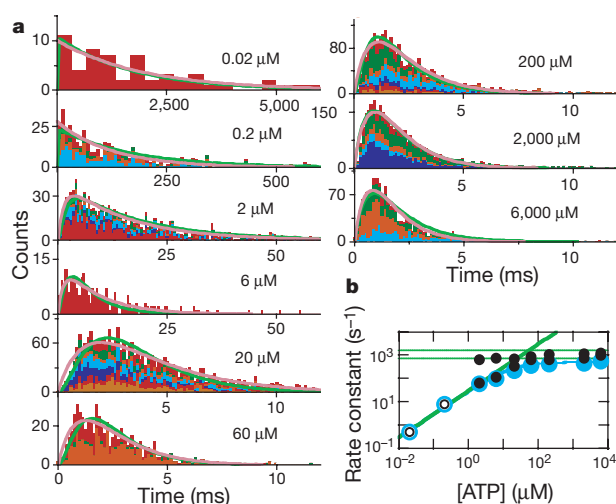
### Two $\sim 1$ -ms reactions before a next step

Figure 8a shows histograms of dwell times between two main steps ( $90^\circ$  or  $120^\circ$ ) that were easily discerned at all [ATP]. At [ATP]  $\ll K_m$  (0.02 and 0.2  $\mu$ M), the histograms were fitted with a single exponential (pink lines) with a rate proportional to [ATP] (Fig. 8b; open circles), indicating that ATP binding alone sets the pace of rotation at these [ATP]. At 2  $\mu$ M ATP, the histogram starts at the origin at zero dwell and shows a distinct peak, indicating the appearance of another rate-limiting reaction. Between 2 and 60  $\mu$ M, the histograms are explained, roughly, by the ATP-binding reaction and an ATP-independent,  $\sim 0.5\text{-ms}^{-1}$  reaction (Fig. 8b). At [ATP]  $\gg K_m = 15 \mu\text{M}$ , the histograms still show a distinct peak, indicating the presence of at least two reactions. A fit with two rate constants indicated both to be  $\sim 1$  ms (Fig. 8b; filled black circles), which, taken together, can account for the  $0.5\text{-ms}^{-1}$  reaction at intermediate [ATP]. Thus, at least three rate constants, two for the  $\sim 1$ -ms reactions and one for ATP binding, are required to describe the rotation kinetics at all [ATP]. Global fit to all histograms (green lines) showed  $k_{on}^{ATP}$  to be  $(3.0 \pm 0.1) \times 10^7 \text{ M}^{-1} \text{ s}^{-1}$  (consistent with the estimate from Fig. 3 (above) and previous values in actin<sup>11</sup> and single-fluorophore<sup>18</sup> assays), and the other two rates to be  $1.64 \pm 0.06 \text{ ms}^{-1}$  and  $0.71 \pm 0.02 \text{ ms}^{-1}$ . Of the latter two reactions, we cannot discriminate which is first, and the two rates do not differ significantly (see also Fig. 8b).

Thus we propose in Fig. 7a that two  $\sim 1$ -ms dwells separate the  $90^\circ$  and  $30^\circ$  substeps. A simple explanation is that the first dwell is terminated by release of a hydrolysis product (phosphate or ADP) and the second by release of the other product. Alternatively, the first reaction may be splitting of ATP to ADP and phosphate, and the second reaction the release of the two. Either interpretation is



**Figure 7** Proposed mechanism for  $F_1$  rotation. **a**, Rotation scheme. ATP with asterisk represents ATP or ADP + phosphate; ADP (alone) may be phosphate or ADP + phosphate. **b**, Stepping time courses expected from **a**. **c**, Highly schematic diagram for the potential energy  $\Psi$  for  $\gamma$ -subunit rotation. Each coloured line shows  $\Psi$  in one of the four states in **a**. The orientation of the  $\gamma$ -subunit in state A (in **a**) is taken as  $0^\circ$ .



**Figure 8** Dwells between main steps. **a**, Histograms of dwell times between two main (90° or 120°) steps at various [ATP]. Total counts in each histogram are 60, 463, 1,145, 2,862, 631, 2,384, 3,262 and 1,457 in the order of 0.02–6,000 μM. Histograms for individual runs are distinguished by colours; they are added to constitute a whole histogram. Pink lines at 0.02 and 0.2 μM ATP are single-exponential fits, constant  $\exp(-k\delta)$ , with  $k$  shown in open circles in **b**. Pink lines at other [ATP] are fits with two rate constants, constant  $\exp(-k_1\delta - \exp(-k_2\delta))$ , with  $k_1$  and  $k_2$  shown (filled black circles in **b**). Green lines show the result of a global fit to the individual histograms ( $n=38$ ; equal weight for each count) with sequential reactions (Fig. 7a) starting with ATP binding at the rate  $k_{\text{on}}^{\text{ATP}}[\text{ATP}]$  and two ATP-independent reactions with rates  $k_1$  and  $k_2$ .  $k_{\text{on}}^{\text{ATP}} = (3.0 \pm 0.1) \times 10^7 \text{ M}^{-1} \text{ s}^{-1}$ ,  $k_1 = 1.64 \pm 0.06 \text{ ms}^{-1}$ , and  $k_2 = 0.71 \pm 0.02 \text{ ms}^{-1}$  (s.e.m.). **b**, ATP dependence of the rate constants. Blue circles show the total rate,  $k$  or  $k_1 k_2 / (k_1 + k_2)$  for the individual fits. Green lines show the rate constants obtained in the global analysis, and the total rate,  $[k_1^{-1} + k_2^{-1} + (k_{\text{on}}^{\text{ATP}}[\text{ATP}])^{-1}]^{-1}$ , is shown by the blue line.

consistent with the biochemical evidence that the releases of ADP and phosphate occur at similar rates<sup>21</sup>.

## Discussion

In our rotation scheme (Fig. 7a) one or two of the three catalytic sites are filled at any time with ATP or its product(s) of hydrolysis. This is the so-called bi-site mechanism<sup>2,3,17</sup>, which is the norm at least at submicromolar [ATP]. Previously we have demonstrated rotation at [ATP] as low as 20 nM, indicating that bi-site hydrolysis accompanies rotation<sup>11</sup> and that bi-site is the fundamental mode of rotation<sup>17</sup>. Present results suggest bi-site to be the norm also at physiological (mM) [ATP]. Alternation between two-filled and three-filled states (the tri-site mechanism) has been proposed for hydrolysis at high [ATP], from non-Michaelis–Menten kinetics<sup>22,23</sup> as in Fig. 3, and from quenching of tryptophan fluorescence in the active sites<sup>24</sup>. These results indicating the tri-site mechanism may have been influenced by the MgADP inhibition<sup>25</sup>. In contrast, ATP synthesis by ATP synthase is insensitive to the inhibition and seems to proceed by a bi-site mechanism<sup>26</sup>. Our rotation assay focuses only on active  $F_1$ , and is unaffected by the inhibition. The rotation rate was higher than one-third of the hydrolysis rate at all [ATP] (Fig. 3), implying that part of  $F_1$  in solution was not fully active already at several seconds after exposure to ATP. We do not necessarily deny possible occupancy of three sites, but we claim that filling all sites does not significantly accelerate, nor add power to, rotation.

Presence of the substeps indicates the appearance of two metastable structures during rotation, in which the equilibrium positions of the  $\gamma$ -subunit differ by  $\sim 90^\circ$  (Fig. 7a: A versus B or C). If our bi-site interpretation is correct, the crystal structure in Fig. 1a probably corresponds to the two-nucleotide structure B (or C). The other central structure with one nucleotide (A) is yet to be solved. A structure<sup>27</sup> reported recently indicates that the protruding portion

of the  $\gamma$ -subunit, where the bead was attached, is torsionally flexible, therefore it is possible that the substeps revealed here might be an artefact: intrinsic steps of the  $\gamma$ -subunit are always  $120^\circ$ , whereas the bead is somehow obstructed at  $90^\circ$  and lags behind for a few ms. We dismiss this possibility because (1) the bead rotated over  $120^\circ$  within 0.25 ms at 2 mM ATP; (2) the dwell time at  $90^\circ$  was independent of the radius of bead rotation, which presumably reflects differences in bead attachment; and (3) larger beads did not show any sign of obstruction.

In Fig. 7a, binding of ATP drives the  $90^\circ$  substep. By reciprocity, the affinity of the  $\beta$ -subunit for ATP, on the left of the arrow on the central  $\gamma$ -subunit, must increase as the  $\gamma$ -subunit rotates<sup>6,16,17,28</sup>. Likewise, because release of ADP (or phosphate, or both) drives the  $30^\circ$  substep, the affinity of the  $\beta$ -subunit for ADP, on the right of the arrow, must decrease as the  $\gamma$ -subunit rotates over the last  $30^\circ$ . The magnitude of these affinity changes can be estimated as below. We have shown that, at least under a high load,  $F_1$  does 80–90 pN nm of mechanical work per  $120^\circ$  step<sup>11</sup> and that the torque it produces is nearly independent of the rotation angle<sup>16,17</sup> (the potential energy  $\Psi$  for  $\gamma$ -subunit rotation is linearly downhill). Rotation of 291-nm beads gave similar results (R.Y., unpublished observations). These results, combined with Fig. 7a, suggest the diagram in Fig. 7c for  $\Psi$ : in state A in Fig. 7a,  $\Psi_A$  is minimal at  $0^\circ$ ; ATP binding produces state B where  $\Psi_B$  is linearly downhill toward its minimum at  $\sim 90^\circ$ ;  $\Psi_C$  is also minimal at  $\sim 90^\circ$ ; product release recovers state A' where  $\Psi_{A'}$  is again linearly downhill toward  $120^\circ$ . Solid lines have the constant slope of 80 pN nm of work per  $120^\circ$  as indicated by experiment, whereas dashed lines are drawn arbitrarily to provide minima at experimental positions. Affinity for ATP is proportional to  $\exp[(\Psi_A - \Psi_B)/k_B T]$ , where  $k_B T \sim 4.1$  pN nm is the thermal energy at room temperature<sup>6,28</sup>. Thus, rotation from  $0^\circ$  to  $90^\circ$  accompanies an increase in the affinity of more than  $\exp[(60 \text{ pN nm})/k_B T] \sim 2 \times 10^6$ . The affinity for ADP,  $\propto \exp[(\Psi_{A'} - \Psi_C)/k_B T]$ , decreases more than  $\exp[(20 \text{ pN nm})/k_B T] \sim 10^2$ .

These affinity changes account for ATP synthesis by forced clockwise rotation of the  $\gamma$ -subunit. Starting from state A' in Fig. 7a, the affinity for ADP of the empty  $\beta$ -subunit on the right of the arrow increases as the  $\gamma$ -subunit rotates clockwise, and this  $\beta$ -subunit will pick up ADP from the medium. Further rotation decreases the affinity for ATP of the  $\beta$ -subunit carrying the previously synthesized ATP (on the left of the arrow), and this ATP will eventually be released. In this simple scheme, it is the central, asymmetrical  $\gamma$ -subunit that dictates which of the three  $\beta$ -subunits should change its affinity, and to what extent. A corollary of the 'γ-dictator model' is that tri-site operation may take place during hydrolysis if ADP release is somehow slowed down, for example by partial inhibition. After a  $90^\circ$  substep, the  $\gamma$ -subunit is already pointing close to the next  $\beta$ -subunit and signals this  $\beta$ -subunit to bind ATP. ATP binding may therefore occur without waiting for the  $30^\circ$  substep accompanying ADP release; however, this would be an inefficient tri-site mechanism with no additional power and speed, as we claim above.

Synthesis (or hydrolysis) of ATP on a  $\beta$ -subunit does not require much energy, because ATP and ADP + phosphate are in equilibrium on the  $\beta$ -subunit<sup>21,26</sup>. This is in accord with our observation that the hydrolysis reaction, which we presume to occur during the two  $\sim 1$ -ms reactions, is mechanically silent. The principal function of hydrolysis is to allow the release of bound ATP by converting it to products, thereby resetting the machine for the next round of stepping. Complete mechanical silence, however, is unfavourable for efficient ATP synthesis, because clockwise rotation should release ATP, not ADP + phosphate. Indeed, it has been indicated that the equilibrium shifts toward ATP during synthesis<sup>26</sup>. One such mechanism (a switch-less model for  $F_1$  motor<sup>16,17</sup>) is suggested in Fig. 7c, where the minimum in  $\Psi_C$  (after hydrolysis) is placed slightly to the right of the minimum in  $\Psi_B$  (before hydrolysis). Such a slight shift would be undetectable at the present resolution.



Our work is essentially a consolidation and embodiment of the binding-change model<sup>1–3</sup>. Similar to F<sub>1</sub>-ATPase, myosin hosts ATP and its hydrolysis product at near equilibrium<sup>29,30</sup>. This may be the general tactics adopted by many ATP-dependent molecular machines. If so, much of the free-energy drop accompanying ATP hydrolysis occurs in the ATP-binding step<sup>17,29,30</sup>. ATP binding may be the principal source of power in these molecular machines. Indeed, ATP binding induces large conformational changes in molecules such as myosin<sup>31</sup>, kinesin<sup>32,33</sup> and chaperonin<sup>34</sup>. □

## Methods

### Proteins

A mutant ( $\alpha$ -C193S,  $\beta$ -His<sub>10</sub> at amino terminus,  $\gamma$ -S107C,  $\gamma$ -I210C)  $\alpha_3\beta_3\gamma$  subcomplex (referred to here as F<sub>1</sub>) derived from a thermophilic *Bacillus* PS3 was biotinylated at two cysteines ( $\gamma$ -107C and  $\gamma$ -210C) by incubation with fourfold molar excess of 6-N'-[2-(N-maleimido)ethyl]-N-piperazylamidoethyl-D-biotinamide for 1 h at 23 °C. Unbound biotin was removed with PD10 (Amersham Pharmacia). Two biotin moieties per protein were found in an assay using 4-hydroxyazobenzene-2-carboxylic acid<sup>35</sup>. Eight molar excess of streptavidin (Pierce) was added to the biotinylated F<sub>1</sub> and purified on Superdex-200HR (Amersham Pharmacia).

### Beads

Colloidal gold (diameter 40 nm; British BioCell International) was coated with biotinylated BSA by incubating 0.2% colloidal gold in 2 mM potassium phosphate pH 7.0 with 0.4 mg ml<sup>-1</sup> BSA and 0.2 mg ml<sup>-1</sup> sulphosuccinimidyl-N'-[N'-(D-biotinyl)-6-aminohexanoyl]-6'-aminohexanoate (biotin-(AC<sub>5</sub>)<sub>2</sub> sulpo-OSu; Dojin) for 1 h at 23 °C. We stored modified gold particles in a solution containing 2 mM potassium phosphate and 0.05% polyethylene glycol. Polystyrene beads were biotinylated as follows: 2.7% amino beads (108 nm; Polyscience) in 20 mM potassium phosphate pH 7.0 was incubated with 1 mg ml<sup>-1</sup> of biotin-(AC<sub>5</sub>)<sub>2</sub> sulpo-OSu for 1 h at 23 °C. We incubated 2.5% of carboxy beads (196 or 291 nm; Bangs) in 20 mM potassium phosphate with 3.6 mM 5-(((N-(biotinoyl)amino)hexanoyl)amino)pentylamine trifluoroacetate salt (Molecular Probes), 1% 1-ethyl-3-(3-dimethylaminopropyl)-carbodiimide hydrochloride and 1% N-hydroxysulphosuccinimide for 1 h at 23 °C.

### Rate of ATP hydrolysis

Nucleotide-depleted F<sub>1</sub> was prepared<sup>18</sup>, and its ATPase activity was determined at 23 °C with an ATP-regenerating system<sup>18,36</sup> containing 1 mM phosphoenolpyruvate, 200  $\mu$ g ml<sup>-1</sup> pyruvate kinase, 100  $\mu$ g ml<sup>-1</sup> lactate dehydrogenase, 0.15 mM NADH, and indicated MgATP in buffer A (50 mM KCl, 2 mM MgCl<sub>2</sub>, 10 mM 3-[N-morpholino]propane-sulfonic acid-KOH, pH 7.0). The initial hydrolysis rate was determined from the slope of absorbance decrease at 340 nm; the average slope was estimated for the period of 2–5 s (2 mM–12  $\mu$ M ATP), 2–7 s (6–2  $\mu$ M) or 2–12 s (0.2–0.06  $\mu$ M) after the start of the reaction. In the presence of LDAO, hydrolysis kinetics showed a lag, and thus the rate was estimated at steady state during 400–500 s (2 mM–60  $\mu$ M), 1,000–1,500 s (20–6  $\mu$ M) or 4,500–5,000 s (2–0.6  $\mu$ M).

### Microscopy

Beads of 196 nm and 291 nm were observed with transmitted light on an Olympus IX-70 microscope. We observed 40-nm and 108-nm beads with laser dark-field microscopy<sup>12</sup> (Fig. 1c). A laser beam (532 nm, diameter 3 mm, 200 mW; Millennia II, Spectra Physics) was introduced into a dark-field condenser (numerical aperture, NA 1.2–1.4; Olympus) to illuminate the specimen obliquely. Light scattered by beads was collected with a  $\times$ 100 objective (NA1.35; Olympus) with its iris diaphragm set to NA  $\sim$ 1.1 to block the direct ray. The field of view was  $\sim$ 50  $\mu$ m. To confirm that we observed single 40-nm beads and not their aggregates, we measured the intensity of beads with regular dark-field microscopy with a halogen lamp, which provided homogeneous illumination. The intensity distribution had a single, large peak and a second small peak at four times the intensity of the first peak. Because an object smaller than the wavelength scatters light in proportion to the square of its volume<sup>37</sup>, these peaks should correspond to single and duplex beads. Most beads were thus single. Twenty-fold reduction in the laser intensity did not affect the rotation speed, indicating that heating by the laser was insignificant.

### Bead rotation assay

A flow cell was made of two KOH-cleaned coverslips separated by two spacers with 50  $\mu$ m thickness<sup>11</sup>. Mixture of 0.1–1 nM beads and F<sub>1</sub> at 10–100 times the bead concentration in buffer B (buffer A plus 5 mg ml<sup>-1</sup> BSA) was applied to the flow cell. Unbound beads were washed out with buffer B, and then buffer B plus MgATP (Mg<sup>2+</sup> 2 mM in excess), 0.1 mg ml<sup>-1</sup> creatine kinase, and 1 mM creatine phosphate was infused. Bead images were recorded as an eight-bit AVI file with a fast framing CCD camera (HiD-Cam, Nac) at 8,000 frames per s at [ATP]  $\geq$  2  $\mu$ M; 125 frames per s at 0.2  $\mu$ M; and 60 frames s<sup>-1</sup> at 0.02  $\mu$ M. The temperature was 23 °C. From each unmodified image, the bead centroid was calculated as  $\sum x_i(I_i - I_{th}) / \sum (I_i - I_{th})$ , where  $x_i$  (or  $y_i$ ) is the pixel coordinate,  $I_i$  the pixel intensity,  $I_{th}$  a threshold value, and the summation was for  $I_i \geq I_{th}$ .

The rotational frictional drag coefficient  $\xi$  for the beads was calculated as follows: for a single bead of radius  $a$  rotating in water with viscosity  $\eta$  ( $= 10^{-9}$  pN nm<sup>-2</sup> s), minimal  $\xi$

is given by  $8\pi\eta a^3$  when the rotation axis is at the bead centre, and maximal  $\xi$  by  $8\pi\eta a^2 + 6\pi\eta a^3 = 14\pi\eta a^2$  when the axis is at a bead edge; for a bead duplex, minimal  $\xi$  is given by  $2 \times 8\pi\eta a^3 = 16\pi\eta a^3$  for a vertical duplex rotating around its centre, whereas maximal  $\xi$  is  $2 \times 8\pi\eta a^2 + 6\pi\eta a^3 + 6\pi\eta a (3a)^2 = 76\pi\eta a^2$  for a horizontal duplex rotating around an edge.

### Analysis of substep kinetics

We superimposed and averaged time courses of individual steps as in Fig. 6a, b. First, we identified all main (90° or 120°) steps in a continuous run by eye. Then, individual steps were aligned on the time axis by positioning the midpoint of each step at time 0 (to within  $\pm$  one frame). Vertical alignment was made by shift, by multiples of 120°. Next, steps other than the central one were eliminated from individual step records: the part earlier than 0.25 ms after the preceding 90° substep (or 120° step when substeps were contiguous) was replaced with a horizontal line at the angle at the 0.25-ms point (for example, a red line in Fig. 6a), and the part later than 0.25 ms before the following 90° substep (or 120° step) was replaced with the angle at that point (Fig. 6a; red line). Finally, we averaged all step records (Fig. 6a, b; thick cyan lines).

Step records averaged over several runs (Fig. 6c; coloured lines) were fitted with theoretical kinetics (Fig. 6c; grey lines) on the basis of the scheme in Fig. 7a; rate constants in the scheme were fixed to the values determined in Fig. 8. For the [ATP]-dependent kinetics for the rise from  $-30^\circ$  to  $0^\circ$ , the grey lines show  $A_{30^\circ} \exp(k_{on}^{ATP} [ATP] t) - A_{30^\circ}$ , where  $A_{30^\circ}$  is the size of the 30° substep,  $k_{on}^{ATP} = 3.0 \times 10^7$  M<sup>-1</sup> s<sup>-1</sup>, and  $t$  ( $< 0$ ) is time from the central 90° substep. For the rise from 90° to 120°, two reactions with rates  $k_1 = 1.64$  ms<sup>-1</sup> and  $k_2 = 0.71$  ms<sup>-1</sup> are assumed, giving  $A_{30^\circ} B \{1 - [k_1 \exp(-k_1 t) - k_2 \exp(-k_2 t)] / (k_2 - k_1)\} + (120^\circ - A_{30^\circ})$  where  $B = \exp(-k_{on}^{ATP} [ATP] \cdot 0.25 \text{ ms})$  accounts for the loss of some of the 30° substeps in the averaging process. Global fit to all experimental curves (allowing a vertical shift for each curve) yielded  $A_{30^\circ} = 29.8 \pm 0.3^\circ$ .

Received 4 October 2000; accepted 26 February 2001.

- Boyer, P. D. & Kohlbrenner, W. in *Energy Coupling in Photosynthesis* (eds Selman, B. R. & Selman-Reimer, S.) 231–240 (Elsevier, Amsterdam, 1981).
- Boyer, P. D. The binding change mechanism for ATP synthase—some probabilities and possibilities. *Biochim. Biophys. Acta* **1140**, 215–250 (1993).
- Boyer, P. D. Catalytic site forms and controls in ATP synthase catalysis. *Biochim. Biophys. Acta* **1458**, 252–262 (2000).
- Cox, G. B., Jans, D. A., Fimmel, A. L., Gibson, F. & Hatch, L. The mechanism of ATP synthase. Conformational change by rotation of the b-subunit. *Biochim. Biophys. Acta* **768**, 201–208 (1984).
- Mitchell, P. Molecular mechanics of protonmotive F<sub>0</sub>F<sub>1</sub> ATPases. Rolling well and turnstile hypothesis. *FEBS Lett.* **182**, 1–7 (1985).
- Oosawa, F. & Hayashi, S. The loose coupling mechanism in molecular machines of living cells. *Adv. Biophys.* **22**, 151–183 (1986).
- Abrahams, J. P., Leslie, A. G. W., Lutter, R. & Walker, J. E. Structure at 2.8 Å resolution of F<sub>1</sub>-ATPase from bovine heart mitochondria. *Nature* **370**, 621–628 (1994).
- Duncan, T. M., Bulgryn, V. V., Zhou, Y., Hutcheon, M. L. & Cross, R. Rotation of subunits during catalysis by *Escherichia coli* F<sub>1</sub>-ATPase. *Proc. Natl Acad. Sci. USA* **92**, 10964–10968 (1995).
- Sabbert, D., Engelbrecht, S. & Junge, W. Intersubunit rotation in active F-ATPase. *Nature* **381**, 623–625 (1996).
- Noji, H., Yasuda, R., Yoshida, M. & Kinoshita, K. Jr Direct observation of the rotation of F<sub>1</sub>-ATPase. *Nature* **386**, 299–302 (1997).
- Yasuda, R., Noji, H., Kinoshita, K. Jr & Yoshida, M. F<sub>1</sub>-ATPase is a highly efficient molecular motor that rotates with discrete 120° steps. *Cell* **93**, 1117–1124 (1998).
- Kudo, S., Magariyama, Y. & Aizawa, S. Abrupt changes in flagellar rotation observed by laser dark-field microscopy. *Nature* **346**, 677–680 (1990).
- Miyata, H. *et al.* Stepwise motion of an actin filament over a small number of heavy meromyosin molecules is revealed in an *in vitro* motility assay. *J. Biochem. (Tokyo)* **115**, 644–647 (1994).
- Weber, P. C., Ohlendorf, D. H., Wendoloski, J. J. & Salemme, F. R. Structural origins of high-affinity biotin binding to streptavidin. *Science* **243**, 85–88 (1989).
- He, X. M. & Carter, D. C. Atomic structure and chemistry of human serum albumin. *Nature* **358**, 209–215 (1992).
- Kinoshita, K. Jr, Yasuda, R. & Noji, H. F<sub>1</sub>-ATPase: a highly efficient rotary ATP machine. *Essays Biochem.* **35**, 3–18 (2000).
- Kinoshita, K. Jr, Yasuda, R., Noji, H. & Adachi, K. A rotary molecular motor that can work at near 100% efficiency. *Phil. Trans. R. Soc. Lond. B* **355**, 473–489 (2000).
- Adachi, K. *et al.* Stepping rotation of F<sub>1</sub>-ATPase visualized through angle-resolved single-fluorophore imaging. *Proc. Natl Acad. Sci. USA* **97**, 7243–7247 (2000).
- Jault, J.-M. *et al.* The  $\alpha_3\beta_3\gamma$  complex of the F<sub>1</sub>-ATPase from the thermophilic *Bacillus* PS3 containing the  $\alpha$ D<sub>261</sub>N substitution fails to dissociate inhibitory MgADP from a catalytic site when ATP binds to noncatalytic sites. *Biochemistry* **34**, 16412–16418 (1995).
- Matsui, T. *et al.* Catalytic activity of the  $\alpha_3\beta_3\gamma$  complex of F<sub>1</sub>-ATPase without noncatalytic nucleotide binding site. *J. Biol. Chem.* **272**, 8215–8221 (1997).
- Cunningham, D. & Cross, R. L. Catalytic site occupancy during ATP hydrolysis by MF<sub>1</sub>-ATPase. Evidence for alternating high affinity sites during steady-state turnover. *J. Biol. Chem.* **263**, 18850–18855 (1988).
- Gresser, M. J., Myers, J. A. & Boyer, P. D. Catalytic site cooperativity of beef heart mitochondrial F<sub>1</sub> adenosine triphosphatase. Correlations of initial velocity, bound intermediate, and oxygen exchange measurements with an alternating three-site model. *J. Biol. Chem.* **257**, 12030–12038 (1982).
- Jault, J.-M. *et al.* The  $\alpha_3\beta_3\gamma$  subcomplex of the F<sub>1</sub>-ATPase from the thermophilic *Bacillus* PS3 with the  $\beta$ T165S substitution does not entrap inhibitory MgADP in a catalytic site during turnover. *J. Biol. Chem.* **271**, 28818–28824 (1996).
- Weber, J., Wilke-Mounts, S., Lee, R. S., Grell, E. & Senior, A. E. Specific placement of tryptophan in the catalytic sites of *Escherichia coli* F<sub>1</sub>-ATPase provides a direct probe of nucleotide binding: maximal ATP hydrolysis occurs with three sites occupied. *J. Biol. Chem.* **268**, 20126–20133 (1993).
- Milgrom, Y. M., Murataliev, M. B. & Boyer, P. D. Bi-site activation occurs with the native and nucleotide-depleted mitochondrial F<sub>1</sub>-ATPase. *Biochem. J.* **330**, 1037–1043 (1998).

26. Zhou, J.-M. & Boyer, P. D. Evidence that energization of the chloroplast ATP synthase favors ATP formation at the tight binding catalytic site and increases the affinity for ADP at another catalytic site. *J. Biol. Chem.* **268**, 1531–1538 (1993).
27. Gibbons, C., Montgomery, M. G., Leslie, A. G. W. & Walker, J. E. The structure of the central stalk in bovine F<sub>1</sub>-ATPase at 2.4 Å resolution. *Nature Struct. Biol.* **7**, 1055–1061 (2000).
28. Wang, H. & Oster, G. Energy transduction in the F<sub>1</sub> motor of ATP synthase. *Nature* **396**, 279–282 (1998).
29. Wolcott, R. G. & Boyer, P. D. The reversal of the myosin and actomyosin ATPase reactions and the free energy of ATP binding to myosin. *Biochem. Biophys. Res. Commun.* **57**, 709–716 (1974).
30. Mannherz, H. G., Schenck, H. & Goody, R. S. Synthesis of ATP from ADP and inorganic phosphate at the myosin-subfragment 1 active site. *Eur. J. Biochem.* **48**, 287–295 (1974).
31. Houdusse, A., Szent-Györgyi, A. G. & Cohen, C. Three conformational states of scallop myosin S1. *Proc. Natl Acad. Sci. USA* **97**, 11238–11243 (2000).
32. Rice, S. *et al.* A structural change in the kinesin motor protein that drives motility. *Nature* **402**, 778–784 (1999).
33. Schnitzer, M. J., Visscher, K. & Block, S. M. Force production by single kinesin motors. *Nature Cell Biol.* **2**, 718–723 (2000).
34. Sigler, P. B. *et al.* Structure and function in GroEL-mediated protein folding. *Annu. Rev. Biochem.* **67**, 581–608 (1998).
35. Kunioka, Y. & Ando, T. Innocuous labeling of the subfragment-2 region of skeletal muscle heavy meromyosin with a fluorescent polyacrylamide nanobead and visualization of individual heavy meromyosin molecules. *J. Biochem. (Tokyo)* **119**, 1024–1032 (1996).
36. Kato, Y., Sasayama, T., Muneyuki, E. & Yoshida, M. Analysis of time-dependent change of *Escherichia coli* F<sub>1</sub>-ATPase activity and its relationship with apparent negative cooperativity. *Biochim. Biophys. Acta* **1231**, 275–281 (1995).
37. Born, M. & Wolf, E. *Principles of Optics* 7th edn. (Cambridge Univ. Press, Cambridge, 1999).

**Supplementary information** is available on Nature's World-Wide Web site (<http://www.nature.com>).

## Acknowledgements

We thank T. Ariga for sample preparation; A. Kusumi for colloidal gold; T. Hisabori, E. Muneyuki, T. Nishizaka, K. Adachi, C. Gosse, M. Y. Ali, S. Ishiwata and G. W. Feigenson for critical discussions; and H. Umezawa for laboratory management. This work was supported in part by Grants-in-Aid from the Ministry of Education, Science, Sports and Culture of Japan.

Correspondence and requests for materials should be addressed to K.K.

# Efficient mixing at low Reynolds numbers using polymer additives

Alexander Groisman\* & Victor Steinberg\*†

\* Department of Physics of Complex Systems, The Weizmann Institute of Science, 76100 Rehovot, Israel

† Present address: Department of Applied Physics, Caltech, Pasadena, California 91125, USA

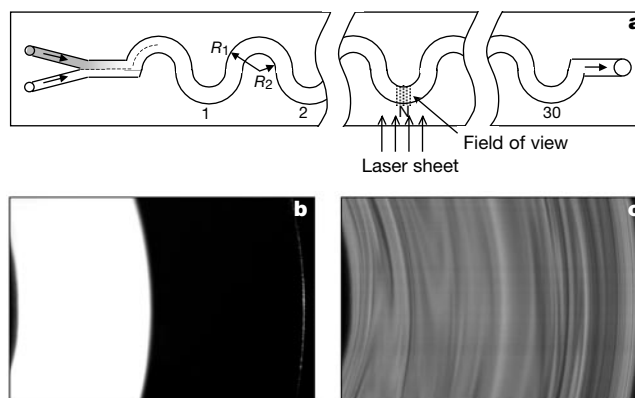
Mixing in fluids is a rapidly developing area in fluid mechanics<sup>1–3</sup>, being an important industrial and environmental problem. The mixing of liquids at low Reynolds numbers is usually quite weak in simple flows, and it requires special devices to be efficient. Recently, the problem of mixing was solved analytically for a simple case of random flow, known as the Batchelor regime<sup>4–8</sup>. Here we demonstrate experimentally that very viscous liquids containing a small amount of high-molecular-weight polymers can be mixed quite efficiently at very low Reynolds numbers, for a simple flow in a curved channel. A polymer concentration of only 0.001% suffices. The presence of the polymers leads to an elastic instability<sup>9</sup> and to irregular flow<sup>10</sup>, with velocity spectra corresponding to the Batchelor regime<sup>4–8</sup>. Our detailed observations of the mixing in this regime enable us to confirm several important theoretical predictions: the probability distributions of the concentration exhibit exponential tails<sup>6,8</sup>, moments of the distribution decay exponentially along the flow<sup>8</sup>, and the spatial correlation function of concentration decays logarithmically.

Solutions of flexible high-molecular-weight polymers are visco-elastic fluids<sup>11</sup>. There are elastic stresses that appear in the polymer solutions in a flow, and that grow nonlinearly with the flow rate<sup>11</sup>. This can lead to many special flow effects, including purely elastic transitions<sup>9,12,13</sup> that qualitatively change the character of the flow at vanishingly small Reynolds numbers,  $Re$ . As a result of such transitions secondary vortical flows appear in different systems, where the primary motion is a curvilinear shear flow. The onset of

those secondary flows depends on the Weissenberg number,  $Wi = \lambda\dot{\gamma}$ , where  $\lambda$  is the polymer relaxation time and  $\dot{\gamma}$  is the shear rate. It plays a role analogous to that of  $Re$  in competition between nonlinearity and dissipation. As has been reported recently, an elastic flow transition can result in a special kind of turbulent motion, elastic turbulence<sup>10</sup>, which arises at arbitrary small  $Re$ . In our experiments, we find that irregular flows excited by the elastic stresses in polymer solutions can lead to quite efficient mixing at very low  $Re$ .

Our mixing experiments were carried out in a curvilinear channel (Fig. 1a) at room temperature,  $22.5 \pm 0.5^\circ\text{C}$ . The total rate of the fluid discharge,  $Q$ , was always kept constant, so that the average time of mixing was proportional to the position,  $N$ , along the channel (Fig. 1a). We used a solution of 65% saccharose and 1% NaCl in water, with viscosity  $\eta_s = 0.153\text{ Pa s}$  and density  $\rho = 1.32\text{ g cm}^{-3}$ , as a solvent for the polymer. We added polyacrylamide, of  $M_w = 1.8 \times 10^7$ , at a concentration of 80 p.p.m. by weight. One of the solutions entering the channel (Fig. 1a) also contained an initial concentration,  $c_0 = 2\text{ p.p.m.}$ , of a fluorescent dye. The solution viscosity was  $\eta = 0.198\text{ Pa s}$  at a shear rate  $\dot{\gamma} = 4\text{ s}^{-1}$ . The relaxation time,  $\lambda$ , estimated from the phase shift between the stress and the shear rate in oscillatory tests, was 1.4 s. An estimate for the diffusion coefficient of the dye is given by that for the saccharose molecules, which is about  $D = 8.5 \times 10^{-7}\text{ cm}^2\text{ s}^{-1}$ . The characteristic shear rate and the Weissenberg number in the flow are estimated as  $\dot{\gamma} = (2Q/d^2)/(d/2) = 4Q/d^3$  and  $Wi = \lambda(4Q/d^3)$ , respectively, where  $d$  is the channel depth.

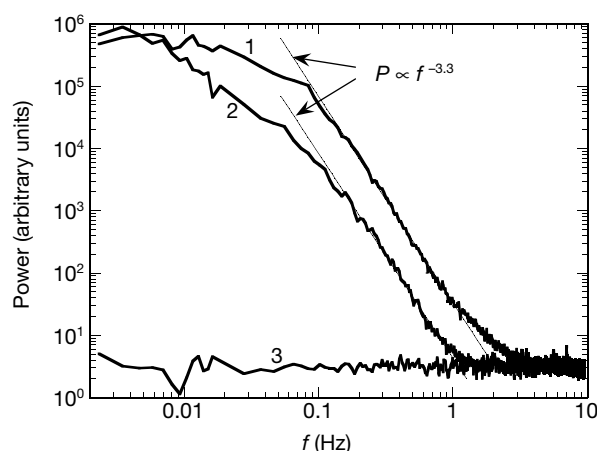
The Reynolds number,  $Re = 2Q\rho/(d\eta)$ , was always quite small. It reached only 0.6 for the highest  $Q$  that we explored. Therefore, flow of the pure solvent remained quite laminar and no mixing occurred; see Fig. 1b. The boundary separating the liquid with and without the dye was smooth and parallel to the direction of the flow, and it only became smeared owing to molecular diffusion as the liquids advanced downstream. The behaviour of the polymer solution was qualitatively different from that of the solvent. The flow was only laminar and stationary up to a value of  $Q$ , corresponding to  $Wi_c = 3.2$  (and  $Re = 0.06$ ), at which point an elastic instability occurred. This instability led to irregular flow and mixing of the



**Figure 1** Experimental set-up and two snapshots of the flow. **a**, Experimental set-up. A channel of depth  $d = 3\text{ mm}$  is machined in a transparent bar of perspex and sealed from above by a transparent window. The channel is square in cross-section, and consists of a sequence of 60 smoothly connected half-rings with inner and outer radii  $R_1 = 3\text{ mm}$  and  $R_2 = 6\text{ mm}$ , respectively. The flow is always observed near the middle of a half-ring on the right-hand side of the bar, when viewed downstream. The half-rings on the right-hand side are numbered starting from the channel entrance, and the number,  $N$ , of the half-ring is a natural linear coordinate along the channel. Two liquids are fed into the channel by two syringe pumps at equal discharge rates through two separate tubes. The two working liquids are always identical; the only difference is that a small amount of a

fluorescent dye (fluorescein) is added to one of them. The channel is illuminated from the side by an argon-ion laser beam converted by two cylindrical lenses to a broad sheet of light, with a thickness of about  $40\text{ }\mu\text{m}$  in the region of observation. The fluorescent light emitted by the liquid in the perpendicular direction is projected onto a charge-coupled device (CCD) camera and digitized by an 8-bit,  $512 \times 512$  frame grabber. The concentration of the dye is evaluated from the intensity of the light, which was found to be proportional to the concentration. **b**, **c**, Snapshots of the flow at  $N = 29$ . Bright regions correspond to high concentrations of the fluorescent dye. Curved boundaries of the channel are seen on the left and on the right. **b**, Pure solvent at  $Re = 0.16$ . **c**, Polymer solution at the same flow rate, corresponding to  $Wi = 6.7$ .





**Figure 2** Power,  $P$ , of fluctuations of velocity in the middle of the channel at  $N = 12$  as a function of frequency,  $f$ . The flow velocity was measured by a laser Doppler anemometer, when the region of the laser beam crossing was made very small,  $15 \times 15 \times 40 \mu\text{m}$ , to decrease the gradient noise. The spectra of the velocity components for the polymer solution are shown: along the mean flow (curve 1), across the mean flow (curve 2) and for the pure solvent across the mean flow (curve 3). The mean velocity for the polymer solution was  $\bar{V} = 6.6 \text{ mm s}^{-1}$ . The root mean square, r.m.s., of the fluctuations,  $V_{\text{r.m.s.}}$ , was  $0.09\bar{V}$  and  $0.04\bar{V}$  in the longitudinal and transverse directions, respectively.

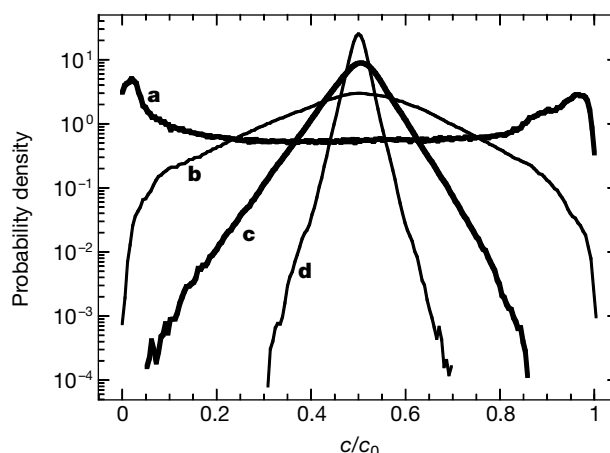
liquids; see Fig. 1c. The experiments were carried out at  $Q$  values of about double the value at the onset of the flow instability,  $Wi = 6.7$ , where homogeneity of the mixture at the exit of the channel was the highest.

Power spectra of fluctuations of velocity in the centre of the channel at  $Wi = 6.7$  are shown in Fig. 2. The spectra of both longitudinal and transversal velocity components do not exhibit any distinct peaks and have broad regions of a power-law decay that is typical for turbulent flow. They resemble the power spectra found at similar  $Wi$  values in flow of the same polymer solution between two parallel plates in the regime of elastic turbulence<sup>10</sup>. So, we believe the origin of the irregular flow is the same here as in ref. 10.

According to the Taylor hypothesis fluctuations of the flow velocity in time are mainly due to fluctuations in space advected by the mean flow<sup>2,3</sup>. So the spectra in Fig. 2 imply that the power of the velocity fluctuations scales with the  $k$ -number in space as  $P \propto k^{-3.3}$ . Fluctuations of the velocity gradients should scale then as  $k^{-1.3}$ , so that the flow becomes increasingly homogeneous at small scales, and mixing is mainly due to the largest eddies, which have the size of the whole system<sup>6</sup>. Such flow conditions correspond to the Batchelor regime of mixing<sup>4,5</sup>. It is one of the two simple cases of random flow, where the problem of mixing has been solved analytically<sup>2,5-8</sup>. The mixing in the irregular flow in the channel gave us an opportunity to study mixing in the Batchelor regime in detail and to test a few important theoretical predictions for the first time, to our knowledge.

Mixing of the polymer solution was a random process. Thus it was appropriate to characterize it statistically, using a probability distribution function (PDF) to find different concentrations,  $c$ , of the dye at a point, and the values of the moments,  $M_i$ , of the distribution. An  $i$ th moment is defined as an average,  $\langle (c - \bar{c})^i \rangle / \bar{c}^i$ , where  $\bar{c}$  is the average concentration of the dye, which in our case is equal to  $c_0/2$ . Small values of  $M_i$  mean high homogeneity and good mixing of the liquids. The values of  $M_i$  are all unity at the channel entrance and become zero for perfectly mixed liquids.

Probability distribution functions of the concentrations at  $N = 8$  and  $N = 24$  (with  $M_1 = 0.72$  and  $0.25$ , respectively) are shown in Fig. 3a and b. Dependencies of  $M_1$  and  $M_2$  on the position  $N$  along the channel are presented in Fig. 4. Advanced stages of mixing



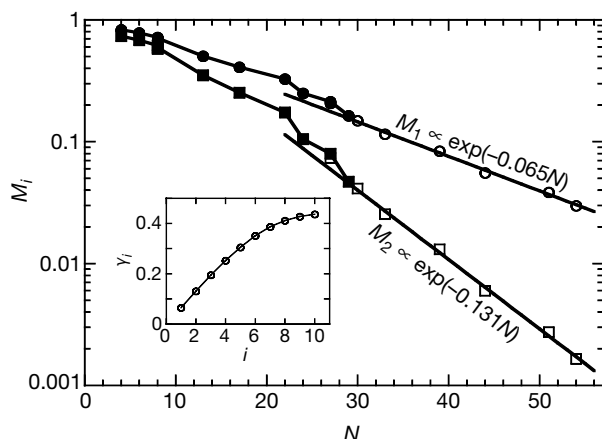
**Figure 3** Plots of probability density function (PDF) of the concentration of the fluorescent dye at different positions. The brightness profile was taken 12.5 times per second along a single line across the channel near the middle of a half-ring (a horizontal line in the middle in Fig. 1b and c). Each plot represents statistics over about  $10^7$  points, corresponding to about 25,000 different moments of time, and a total liquid discharge of  $2 \times 10^3 d^3$ . The regions near the walls of the channel with widths of  $0.1d$  were excluded from the statistics. Curves **a**, **b**: PDF at  $N = 8$  and  $N = 24$ ; curves **c**, **d**: with preliminary mixing, at the positions corresponding to  $N = 39$  and  $N = 54$ , respectively.

corresponding to  $N > 30$  were studied when the liquids were premixed before they entered the channel (Fig. 4). Representative PDFs in the flow with preliminary mixing taken at the positions where  $M_1 = 0.082$  and  $0.030$  are shown in Fig. 3c and d.

Representative spatial autocorrelation functions for the dye concentration are shown in Fig. 5. At the beginning the concentration distribution is strongly influenced by the initial conditions. There are spatially extended uniform regions with maximal and zero dye concentration, so the concentration remains highly correlated over rather large distances (Fig. 5a) and the PDF has maxima near  $c_0$  and zero (Fig. 3a). As the liquid advances downstream, it becomes increasingly homogeneous. Mixing patterns exhibit many fine-scale structures of different brightnesses (Fig. 1c), the correlation distances become shorter (Fig. 5b) and the PDFs of the dye concentration become narrower (Fig. 3b). The PDF in Fig. 3b has a single peak at  $\bar{c}$  and long tails that decay exponentially and touch the limits of the concentration: zero and  $c_0$ .

Further downstream, mixing patterns have characteristic features at similar spatial scales (Fig. 5c–e), but are much more faded. The PDF in Fig. 3c is much narrower than in Fig. 3b and has quite clear exponential tails, in agreement with the theoretical predictions<sup>6,7</sup>, which implies significant intermittency in mixing<sup>2</sup>. The distribution is well confined in a region far from the limits of zero and  $c_0$ . Thus the dependence on the initial conditions should be quite minor by that point. At the last point ( $N = 54$ ,  $M_1 = 0.030$ ) the non-homogeneity in the concentration is hardly seen, and the PDF (Fig. 3d) is very narrow. At an  $N$  of about 30 the autocorrelation function reaches some asymptotic form, which does not change as the liquid moves further downstream. This may correspond to the situation when the small-scale structures are smeared by molecular diffusion at the same rate as they are created by the flow. The autocorrelation functions (Fig. 5c–e) decay logarithmically at the distances above the diffusion length; this is in agreement with the theoretical predictions for the Batchelor regime<sup>4-6</sup>. The spatial structure of concentration fluctuations in the Batchelor regime has been experimentally studied before in different types of flows<sup>3,14,15</sup>. However, agreement between experiment and theory remained rather controversial (see refs 2, 14 and 15 for discussion).

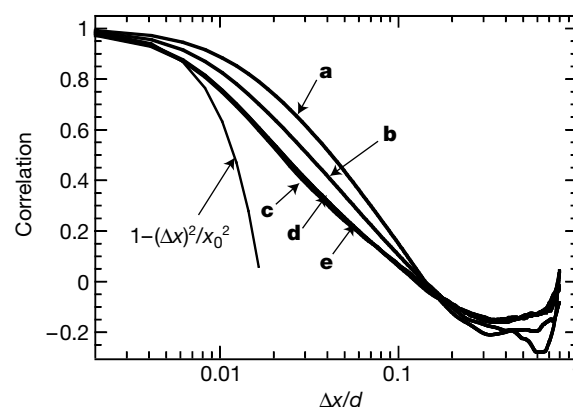
The dependences of  $M_1$  and  $M_2$  on  $N$  in Fig. 4 have inflections at



**Figure 4** Dependence of  $M_1$  (circles) and  $M_2$  (squares) on the position,  $N$ , along the channel. The average flow time is connected to  $N$  as  $\bar{t} = N \times 7.8$  s. In order to observe stages of mixing corresponding to  $N > 30$ , we carried out a series of experiments, where the liquids were premixed before they entered the channel. A shorter channel of the same shape was designed for this purpose and put before the entrance to the original one. As a result of this preliminary mixing, the PDF of the dye concentrations at  $N = 2$  was almost identical to the PDF at  $N = 27$  without the premixing. Filled symbols represent  $M_1$  and  $M_2$  without preliminary mixing. Empty symbols are for the flow with preliminary mixing, when a value of 25 is added to the actual position, in order to match the entrance conditions. We note that these curves are indeed continuations of the dependencies obtained for  $M_1$  and  $M_2$  in the channel without the premixing.

$N$  values between 20 and 30. From the discussion above we suggest that these are due to a transition to an asymptotic regime, where dependence on the initial conditions is lost. The correlation functions at different positions become identical, and PDF of concentrations can be superimposed rather well for  $\Delta c$  of about  $3M_1\bar{c}$  around  $\bar{c}$  by appropriate stretching of the axes. There is no self-similarity in the shapes of PDF at larger  $\Delta c/(M_1\bar{c})$ , however, and the shape permanently changes with  $N$ . We learn from Fig. 4 that both  $M_1$  and  $M_2$  decay exponentially above  $N = 30$ , the rate of the decay being twice as high for  $M_2$  as for  $M_1$ . In addition, the higher-order moments were found to decay exponentially,  $M \propto \exp(-\gamma_i N)$ , which is quite in agreement with the theoretical predictions<sup>8</sup>. We note here that  $M_2$ , which is the variance of fluctuations of the dye concentration, is often considered as an analogue of the energy of turbulent motion<sup>2</sup>. So its logarithmic derivative,  $d \ln M_2 / dt = -\gamma_2 / 7.8 \text{ s}^{-1}$  (see Fig. 4), is the analogue of the energy dissipation rate. The theory also predicts linear growth of the coefficients,  $\gamma_i$ , with  $i$  at small values and saturation of the growth, when  $i$  becomes large<sup>8</sup>. This form of dependence of  $\gamma_i$  on  $i$  is quite confirmed by our experimental results; see inset in Fig. 4. The deviation of the dependence of  $\gamma_i$  on  $i$  from a straight line at large  $i$  values is quantitative evidence for the lack of self-similarity in the mixing. We can see that all our experimental results agree very well with all the theoretical predictions for the Batchelor regime of mixing that we tested<sup>4–8</sup>.

As is suggested by the decay of  $M_1$  in Fig. 4, the polymer solutions are mixed at a characteristic distance of  $\Delta N \approx 15$ . It corresponds to a total path of about  $140d$  and an average flow time of about 120 s, which is three orders of magnitude smaller than the diffusion time,  $d^2/D$ . Using a more concentrated sugar syrup as a solvent, we prepared a polymer solution with  $D$  about 30% smaller and with viscosity and relaxation time about two times larger than those of the original solution. We studied the distribution of  $c$  at  $N = 29$  in a flow of this solution at the same  $Wi$  value of 6.7 as before. Here the characteristic flow velocities were two times lower.  $Re$  was about four times lower, and the ratio between the flow time and the



**Figure 5** Correlation coefficients for the concentration as functions of the distance  $\Delta x$  across the channel (semilogarithmic coordinates). Curves **a–e** correspond to the same conditions as for the PDFs in Fig. 3a–d, and are calculated using the same data arrays. Curve 5 is for  $N = 29$ . The correlation functions coincide for  $N \geq 29$ . At small  $\Delta x$  they have parabolic scaling with a characteristic length  $x_0 = 0.017d \approx 50 \mu\text{m}$ . It is probably defined by the thickness of the illuminating light sheet (about  $40 \mu\text{m}$ ) and by the molecular diffusion scale,  $(D/(V_{rms}/d))^{1/2} \approx 25 \mu\text{m}$ . At larger  $\Delta x$  the scaling is logarithmic.

diffusion time,  $d^2/D$ , was about 1.5 times higher than in the original flow. The measured values of  $M_1$  and  $M_2$  were the same as in Fig. 4, however. Thus the inertial forces and the molecular diffusion did not have any apparent influence on the mixing efficiency. The dependence of the efficiency of mixing at the optimal flow conditions on the concentration of the polymers was surprisingly weak (although  $Wi_c$  increased quickly when the polymer concentration was decreasing). So, for a solution with a polymer concentration of 10 p.p.m. ( $\eta/\eta_s = 1.03$ ),  $M_1$  values as low as 0.22 were reached at  $N = 29$  (and at  $Re = 0.065$ ). Mixing was observed down to a polymer concentration of 7 p.p.m. Thus, very viscous liquids can be efficiently mixed in curvilinear channels at very low flow rates with the aid of polymer additives at very low concentrations. This method of mixing, we believe, may find some industrial and laboratory applications. □

Received 17 October 2000; accepted 5 March 2001.

- Sreenivasan, K. R., Ramshankar, R. & Meneveau, C. Mixing, entrainment, and fractal dimension of interfaces in turbulent flows. *Proc. R. Soc. Lond.* **421**, 79–108 (1989).
- Shraiman, B. I. & Siggia, E. D. Scalar turbulence. *Nature* **405**, 639–645 (2000).
- Warhaft, Z. Passive scalars in turbulent flows. *Annu. Rev. Fluid Mech.* **32**, 203–240 (2000).
- Batchelor, G. K. Small scale variation of convected quantities like temperature in turbulent fluid. *J. Fluid Mech.* **5**, 113–133 (1959).
- Kraichnan, R. H. Convection of a passive scalar by a quasi-uniform random straining field. *J. Fluid Mech.* **64**, 737–762 (1974).
- Chertkov, M., Falkovich, G., Kolokolov, I. & Lebedev, V. Statistics of a passive scalar advected by a large scale 2-dimensional velocity field—analytic solution. *Phys. Rev. E* **51**, 5609–5627 (1995).
- Shraiman, B. I. & Siggia, E. D. Lagrangian path integrals and fluctuations in random flow. *Phys. Rev. E* **49**, 2912–2927 (1994).
- Balkovsky, E. & Fouxon, A. Universal long-time properties of Lagrangian statistics in the Batchelor regime and their application to the passive scalar problem. *Phys. Rev. E* **60**, 4164–4174 (1999).
- Larson, R. G., Shaqfeh, E. S. G. & Muller, S. J. A purely viscoelastic instability in Taylor-Couette flow. *J. Fluid Mech.* **218**, 573–600 (1990).
- Groisman, A. & Steinberg, V. Elastic turbulence in a polymer solution flow. *Nature* **405**, 53–55 (2000).
- Bird, R. B., Curtiss, C. F., Armstrong, R. C. & Hassager, O. *Dynamics of Polymeric Liquids* (Wiley, New York, 1987).
- Byars, J. A., Öztekin, A., Brown, R. A. & McKinley, G. H. Spiral instabilities in the flow of highly elastic fluids between rotating parallel disks. *J. Fluid Mech.* **271**, 173–218 (1994).
- Joo, J. L. & Shaqfeh, E. S. G. Observations of purely elastic instabilities in the Taylor-Dean flow of a Boger fluid. *J. Fluid Mech.* **262**, 27–73 (1994).
- Miller, P. L. & Dimotakis, P. E. Measurements of scalar power spectra in high Schmidt number turbulent jets. *J. Fluid Mech.* **308**, 129–146 (1996).
- Williams, B. S., Marteau, D. & Gollub, J. P. Mixing of a passive scalar in magnetically forced two-dimensional turbulence. *Phys. Fluids* **9**, 2061–2080 (1997).

## Acknowledgements

We thank G. Falkovich for theoretical guidance and discussions. The work was partially supported by the Minerva Center for Nonlinear Physics of Complex Systems, by a Research Grant from the Henry Gutwirth Fund and by an Israel Science Foundation grant.

Correspondence and requests for materials should be addressed to V.S.  
(e-mail: victor.steinberg@weizmann.ac.il).

## Magnetic-field-induced superconductivity in a two-dimensional organic conductor

S. Uji<sup>\*</sup>, H. Shinagawa<sup>\*</sup>, T. Terashima<sup>\*</sup>, T. Yakabe<sup>\*</sup>, Y. Terai<sup>\*</sup>,  
M. Tokumoto<sup>†</sup>, A. Kobayashi<sup>‡</sup>, H. Tanaka<sup>§</sup> & H. Kobayashi<sup>§</sup>

<sup>\*</sup> National Research Institute for Metals, Tsukuba, Ibaraki 305-0003, Japan

<sup>†</sup> Electrotechnical laboratory, Tsukuba, Ibaraki 305-8568, Japan

<sup>‡</sup> Research Centre for Spectrochemistry, Graduate School of Science, The University of Tokyo, Bunkyo-ku, Tokyo 113-0033, Japan

<sup>§</sup> Institute for Molecular Science, Okazaki, Aichi 444-8585, Japan

The application of a sufficiently strong magnetic field to a superconductor will, in general, destroy the superconducting state. Two mechanisms are responsible for this. The first is the Zeeman effect<sup>1,2</sup>, which breaks apart the paired electrons if they are in a spin-singlet (but not a spin-triplet) state. The second is the so-called 'orbital' effect, whereby the vortices penetrate into the superconductors and the energy gain due to the formation of the paired electrons is lost<sup>3</sup>. For the case of layered, two-dimensional superconductors, such as the high- $T_c$  copper oxides, the orbital effect is reduced when the applied magnetic field is parallel to the conducting layers<sup>4</sup>. Here we report resistance and magnetic-torque experiments on single crystals of the quasi-two-dimensional organic conductor  $\lambda$ -(BETS)<sub>2</sub>FeCl<sub>4</sub>, where BETS is bis(ethylenedithio)tetraselenafulvalene<sup>5–8</sup>. We find that for magnetic fields applied exactly parallel to the conducting layers of the crystals, superconductivity is induced for fields above 17 T at a temperature of 0.1 K. The resulting phase diagram indicates that the transition temperature increases with magnetic field, that is, the superconducting state is further stabilized with magnetic field.

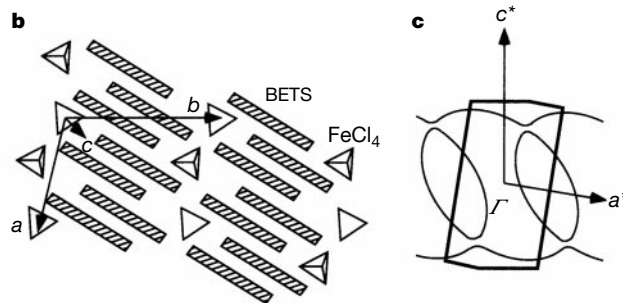
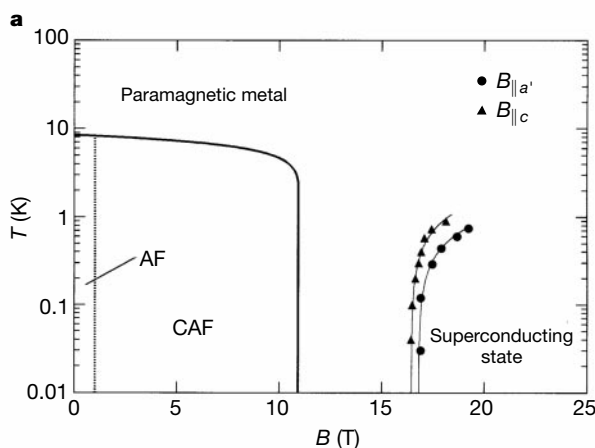
Studies on organic conductors have brought us deep understanding of physics in low-dimensional electronic systems. Members of the BETS family containing magnetic Fe ions among various organic conductors have been extensively studied over the past ten years because strong competition is expected between the antiferromagnetic order of the Fe moments and the superconductivity<sup>5,6</sup>. Of these,  $\lambda$ -(BETS)<sub>2</sub>FeCl<sub>4</sub> has an unusual phase diagram, shown in Fig. 1a. At zero magnetic field,  $\lambda$ -(BETS)<sub>2</sub>FeCl<sub>4</sub> shows a metal–insulator transition at 8 K (ref. 5), whereas the iso-structural salt  $\lambda$ -(BETS)<sub>2</sub>GaCl<sub>4</sub> undergoes a superconducting transition around 6 K (ref. 7). The metal–insulator transition is associated with the antiferromagnetic order of the Fe moments<sup>5,8</sup>. The ordered Fe moments are canted by a magnetic field of about 1 T, but the electronic state remains insulating. The insulating phase is destabilized by magnetic fields above about 10 T, and a paramagnetic metallic state is then recovered.

$\lambda$ -(BETS)<sub>2</sub>FeCl<sub>4</sub> has a triclinic unit cell<sup>6</sup>. The planar BETS molecules are stacked along the  $a$  and  $c$  axes, and consequently form two-dimensional conduction layers (Fig. 1b). The FeCl<sub>4</sub> ion (insulating) layer is intercalated between the BETS layers, which makes

the  $b$  axis the least conducting direction. Because of the short interatomic distance between the BETS and FeCl<sub>4</sub>, finite interactions between the  $\pi$  (conduction) electron on the BETS molecules and the Fe<sup>3+</sup> 3d electrons are expected. The band calculation predicts that  $\lambda$ -(BETS)<sub>2</sub>FeCl<sub>4</sub> has one closed (two-dimensional) and two open Fermi surfaces in the metallic phase as shown in Fig. 1c (refs 5, 6).

The needle-like single crystals of  $\lambda$ -(BETS)<sub>2</sub>FeCl<sub>4</sub>, elongating along the  $c$  axis, were prepared by electrochemical oxidation in an appropriate solvent<sup>6</sup>. The resistance was measured by a conventional four-probe a.c. technique with electric current along the  $b^*$  axis, which is perpendicular to the  $a$ – $c$  plane. Four gold wires (of diameter 10  $\mu$ m) were attached to the sample by gold or carbon paint. The magnetic torque was measured by a simple cantilever technique<sup>7</sup>. The experiments were made with a dilution refrigerator and a 20-T superconducting magnet.

The interlayer resistance with a current  $I$  parallel to the  $b^*$  axis ( $I_{\parallel b^*}$ ) for a magnetic field  $B$  parallel to the  $c$  axis ( $B_{\parallel c}$ ) is presented in Fig. 2a. The insulator–metal transition takes place at around 10.5 T. At 0.04 K, the resistance increases with increasing field after the insulator–metal transition, has a broad maximum around 14 T, and then suddenly decreases by three orders of magnitude. Above 18 T, the detected sample voltage becomes smaller than the noise level, which suggests a superconducting phase transition. As temperature



**Figure 1 a**, Temperature versus magnetic field phase diagram for  $\lambda$ -(BETS)<sub>2</sub>FeCl<sub>4</sub> when the magnetic field is applied parallel to the  $a'$  or  $c$  axis. At zero magnetic field, a metal–insulator transition takes place at around 8 K. In the insulating phase, the Fe moments are antiferromagnetically ordered. AF and CAF indicate the antiferromagnetic and canted antiferromagnetic phases, respectively. The superconducting phase is induced above 17 T. **b**, Schematic picture of the crystal structure of  $\lambda$ -(BETS)<sub>2</sub>FeCl<sub>4</sub>. The  $a$ – $c$  plane is the conduction layer and the  $b$  axis is the least conducting axis. **c**, Calculated Fermi surfaces in the  $k_a$ – $k_c$  plane. The area surrounded by thick lines shows the first Brillouin zone. There are one closed (two-dimensional) and two open Fermi surfaces.



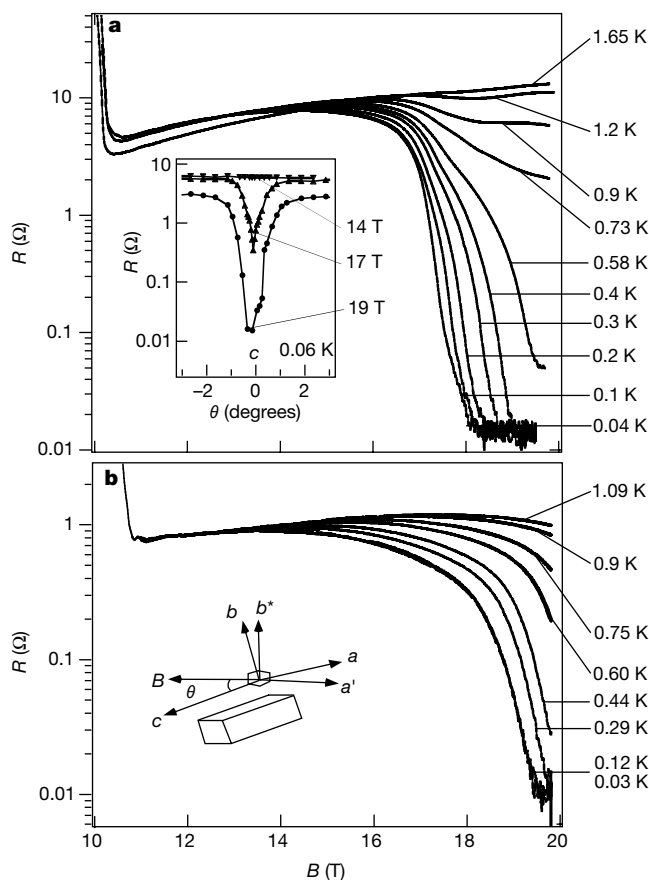
increases, the sudden decrease of the resistance at high fields is suppressed and the resistance shows only monotonic increase at 1.65 K. Hysteresis is not evident at high fields above 16 T, although it is clearly observable near the insulator–metal transition. The inset shows the field-direction dependence of the resistance at three different fields, where the field is rotated in the  $b^*c$  plane. At 14 T, the resistance is constant in this angle range, but shows a significant decrease for  $B_{\parallel c}$  at higher fields, 17 T and 19 T. The low resistance is observed only in a very limited angle range of  $|\theta| < 0.3^\circ$ .

Figure 2b shows the interlayer resistance of a different sample for  $B$  parallel to the  $a'$  ( $B_{\parallel a'}$ ). The  $a'$  axis is defined as that perpendicular to both the  $c$  and  $b^*$  axes. For  $B_{\parallel a'}$ , we observe a sudden decrease of the resistance at relatively higher fields. Compared with the results for  $B_{\parallel c}$ , the resistance seems to decrease more slowly. We obtained a similar sudden decrease in resistance for four different samples. Therefore, we believe that this is intrinsic behaviour for  $\lambda$ -(BETS)<sub>2</sub>FeCl<sub>4</sub>. Judging from the above results, we conclude that the superconductivity is induced only when the magnetic field is parallel to the conduction layers ( $a$ – $c$  plane).

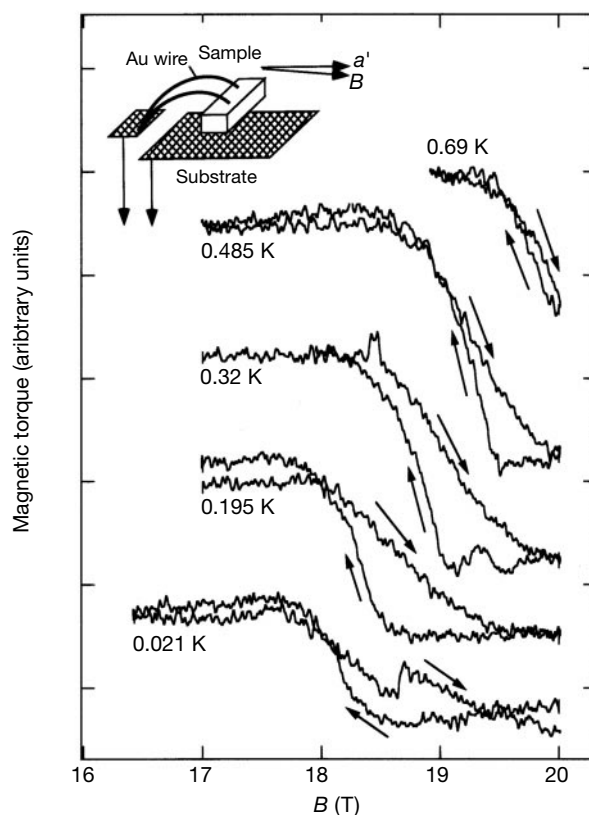
To obtain evidence of the phase transition in thermodynamic quantities under parallel-field conditions, we measured the magnetic torque (Fig. 3). In this experiment, the sample is suspended over a conductive substrate by fine gold wires, forming a capacitor. The rotatable sample holder is mounted under very homogeneous fields (homogeneity of  $10^{-4} \text{ cm}^{-3}$ ), so that the magnetic force arises only from the magnetic torque. The change in the capacitance

between the sample and the substrate, that is, the sample movement, is measured as a function of magnetic field, which directly corresponds to the change in the magnetic torque<sup>9</sup>. To observe the finite change of the magnetic torque, the magnetic field is tilted from the  $a'$  axis by about  $0.5^\circ$ . In the field range where the resistance decreases steeply, we see a significant change in the magnetic torque, showing the presence of a phase transition. At 0.021 K, the torque changes at a field intensity of above 17.5 T. This change shows that the sample is moved towards the substrate by the magnetic torque when  $B > 17.5$  T. The opposite change in the torque is observed when the field is tilted from the  $a'$  axis to the opposite direction. The results verify that the sample turns the  $a'$  axis toward the field direction, that is, the electronic state is more stabilized by the field parallel to the conduction layers in the high-field range. The sample maintains its position in a wider field range during the down sweep so that the field remains between the conduction layers. This behaviour suggests that the magnetic field is pinned between the superconducting layers for the down sweep. The pinning model explains why we observe a large hysteresis in the magnetic torque, but not in the resistance (Fig. 2).

We define the transition field  $B_c$  when  $R/R_N = 0.5$ , where  $R_N$  is the resistance in the metallic state given by extrapolating  $R(B)$  from the low-field range. The transition field defined in this way corresponds to the onset of the magnetic torque change. The obtained phase diagram is presented in Fig. 1a. The transition field is slightly different for each of two field directions. Here we note two interesting features. First, the transition field seems to arise quickly at a certain field. This shows that the paramagnetic metallic state intervenes between the insulating and superconducting phases.



**Figure 2** Interlayer resistance (for current parallel to the  $b^*$  axis) when the magnetic field is exactly parallel to the conduction layers. A steep decrease is evident at high fields, suggesting a superconducting transition. **a**, Resistance when the magnetic field is parallel to the  $c$  axis. The inset shows the field-direction dependence of the resistance under the field in the  $b^*c$  plane. **b**, Resistance when the magnetic field is parallel to the  $a'$  axis.



**Figure 3** Magnetic torque measured by cantilever technique. The data are shifted for clarity. The inset shows the configuration of the sample and the substrate. The magnetic field is tilted from the  $a'$  axis in the  $a'b^*$  plane by about  $0.5^\circ$ . A significant change associated with hysteresis, showing the presence of a phase transition, is observable at high fields.

Second, a positive curvature of the phase boundary occurs towards a high field. The positive curvature suggests that the transition field increases with temperature.

Field-induced superconductivity has been reported for paramagnetic  $\text{Eu}_{1-x}\text{Sn}_x\text{Mo}_6\text{S}_8$  (ref. 10). This material is a superconductor with  $T_c = 3.8\text{ K}$ . As the magnetic field increases, the superconductivity is restored above 4 T and below 0.1 K, after it is destroyed at about 1 T. This phenomenon is well understood in terms of the Jaccarino–Peter compensation effect<sup>11</sup>; the internal magnetic field created by the Eu moments through the exchange interaction is compensated by the external magnetic field. For  $\lambda\text{-(BETS)}_2\text{FeCl}_4$ , because an exchange interaction between the Fe moments and the conduction electrons is expected, the Jaccarino–Peter compensation effect may be a possible mechanism. However, this effect cannot explain how a very small amount of the magnetic field (at about 0.1 T) along the  $b^*$  axis destroys the superconducting state, because the paramagnetic Fe moments are aligned along the external field irrespective of the field direction. The surprisingly strong anisotropic superconducting transition for  $\lambda\text{-(BETS)}_2\text{FeCl}_4$  suggests that the low dimensionality of the electronic system is closely related to the mechanism of the superconductivity.

For quasi-two-dimensional electronic systems such as  $\lambda\text{-(BETS)}_2\text{FeCl}_4$ , when a high magnetic field is applied exactly parallel to the conduction layers, the motion of the electrons is confined onto a single conduction layer. This is field-induced dimensional crossover from quasi-two-dimensional to two-dimensional. In this case, the orbital effect, one of the mechanisms destroying the superconductivity, is largely suppressed. However, if even a small amount of the magnetic field is perpendicular to the conduction layers, the electrons drift in the interlayer direction and the orbital effect is restored. This picture of dimensional crossover seems consistent with the experimental observation that very low magnetic fields along the  $b^*$  axis destroy superconductivity (Fig. 2a). On the basis of the dimensional crossover, various theoretical models predicting the existence of high-field superconducting phases have been proposed<sup>12–16</sup>. However, all the theories tacitly assume that the superconductivity is stable at zero field.

The superconductivity of the iso-structural salt  $\lambda\text{-(BETS)}_2\text{GaCl}_4$  ( $T_c = 6\text{ K}$ ) survives in a magnetic field of up to 5 T that is perpendicular to the conduction layers<sup>17</sup>. This field is much larger than that for  $\lambda\text{-(BETS)}_2\text{FeCl}_4$  (at about 0.1 T). We have carefully measured the resistance for  $\lambda\text{-(BETS)}_2\text{GaCl}_4$  under fields exactly parallel to the conducting layers, but superconductivity was not restored at fields up to 20 T after the superconductivity was destroyed at around 13 T. The completely different phase diagrams of  $\lambda\text{-(BETS)}_2\text{GaCl}_4$  and  $\lambda\text{-(BETS)}_2\text{FeCl}_4$  suggest that the interaction between the Fe moments and the conduction electrons, as well as the low dimensionality of the electronic system, strongly affects the emergence of the superconductivity. It seems likely that the pairing interaction between the  $\pi$  electrons on the BETS molecules arises from the magnetic fluctuation through the paramagnetic Fe moments. □

Received 6 December 2000; accepted 14 February 2001.

- Chandrasekhar, B. S. A note on the maximum critical field of high-field superconductors. *Appl. Phys. Lett.* **1**, 7–8 (1962).
- Clogston, A. M. Upper limit for the critical field in hard superconductors. *Phys. Rev. Lett.* **15**, 266–267 (1962).
- Tinkham, M. *Introduction to Superconductivity* (McGraw-Hill, New York, 1975).
- Klemm, R. A., Luther, A. & Beasley, M. R. Theory of the upper critical field in layered superconductors. *Phys. Rev. B* **12**, 877–891 (1975).
- Brossard, L. et al. Interplay between chains of  $S = 5/2$  localised spins and two-dimensional sheets of organic donors in the synthetically built magnetic multilayer  $\lambda\text{-(BETS)}_2\text{FeCl}_4$ . *Eur. Phys. J. B* **1**, 439–452 (1998).
- Kobayashi, H. et al. New BETS conductors with magnetic anions (BETS=bis(ethylenedithio)tetraselenafulvalene). *J. Am. Chem. Soc.* **118**, 368–377 (1996).
- Kobayashi, H. et al. Electric and magnetic properties and phase diagram of a series of organic superconductors  $\lambda\text{-(BETS)}_2\text{GaX}_2\text{Y}_{1-x}$  [X, Y=Fe, Cl, Br;  $0 < x < 2$ ]. *Phys. Rev. B* **56**, R8526–R8529 (1997).

- Akutsu, H. et al. Highly correlated organic conductor with magnetic anions exhibiting a  $\pi$ -d coupled metal-insulator transition,  $\lambda\text{-(BETS)}_2\text{FeBr}_2\text{Cl}_{1-x}$ . *J. Am. Chem. Soc.* **119**, 12681–12682 (1997).
- Brooks, J. S., Naughton, M. J., Ma, Y. P., Chaikin, P. M. & Chamberlin, R. V. Small sample magnetometers for simultaneous magnetic and resistive measurements at low temperatures and high magnetic fields. *Rev. Sci. Instrum.* **58**, 117–121 (1987).
- Meul, H. W. et al. Observation of magnetic-field-induced superconductivity. *Phys. Rev. Lett.* **53**, 497–500 (1984).
- Jaccarino, V. & Peter, M. Ultra-high-field superconductivity. *Phys. Rev. Lett.* **9**, 290–292 (1962).
- Lebed, A. G. Reversible nature of the orbital mechanism for the suppression of superconductivity. *JETP Lett* **44**, 114–117 (1986).
- Dupuis, N., Montambaux, G. & Sa de Melo, C. A. R. Quasi-one-dimensional superconductors in strong magnetic field. *Phys. Rev. Lett.* **70**, 2613–2616 (1993).
- Miyazaki, M., Kishigi, K. & Hasegawa, Y. Superconductivity of quasi-one- and quasi-two-dimensional tight binding electrons in a magnetic field. *J. Phys. Soc. Jpn* **67**, 2618–2621 (1998).
- Shimahara, H. Coexistence of singlet and triplet attractive channels in the pairing interactions mediated by antiferromagnetic fluctuations. *J. Phys. Soc. Jpn* **69**, 1966–1969 (2000).
- Sa de Melo, C. A. R. Paramagnetism and reentrant behaviour in quasi-one-dimensional superconductors at high magnetic fields. *Physica C* **260**, 224–230 (1996).
- Tanatar, M. A. et al. Anisotropy of the upper critical field of the organic conductor  $\lambda\text{-(BETS)}_2\text{GaCl}_4$ . *J. Supercond.* **12**, 511–514 (1999).

Correspondence and requests for materials should be addressed to S.U. (e-mail: uji@nrim.go.jp).

## Solid acids as fuel cell electrolytes

Sossina M. Haile, Dane A. Boysen, Calum R. I. Chisholm & Ryan B. Merle

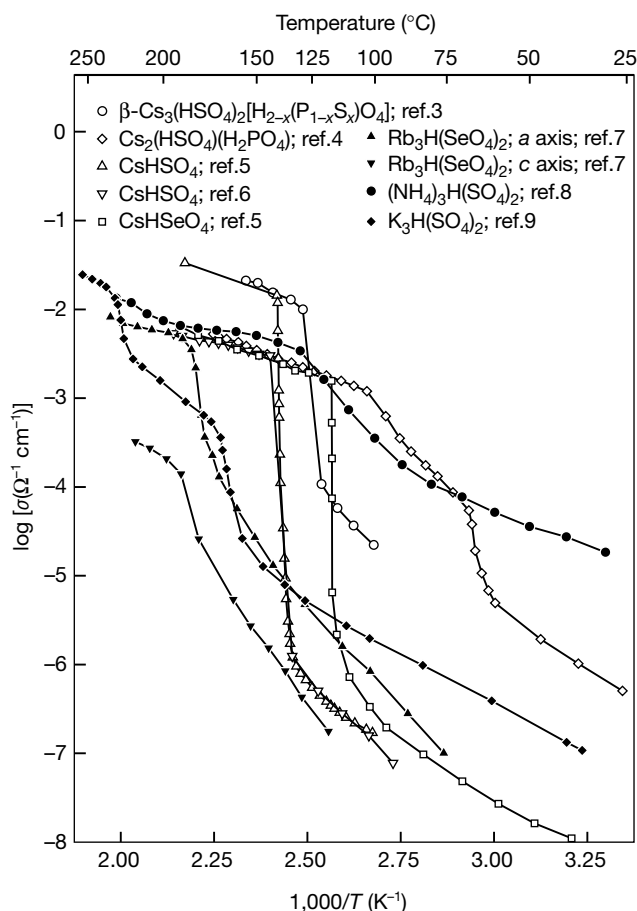
Materials Science, California Institute of Technology, Pasadena, California 91125, USA

Fuel cells are attractive alternatives to combustion engines for electrical power generation because of their very high efficiencies and low pollution levels. Polymer electrolyte membrane fuel cells are generally considered to be the most viable approach for mobile applications. However, these membranes require humid operating conditions, which limit the temperature of operation to less than 100 °C; they are also permeable to methanol and hydrogen, which lowers fuel efficiency. Solid, inorganic, acid compounds (or simply, solid acids) such as  $\text{CsHSO}_4$  and  $\text{Rb}_3\text{H}(\text{SeO}_4)_2$  have been widely studied because of their high proton conductivities and phase-transition behaviour. For fuel-cell applications they offer the advantages of anhydrous proton transport and high-temperature stability (up to 250 °C). Until now, however, solid acids have not been considered viable fuel-cell electrolyte alternatives owing to their solubility in water and extreme ductility at raised temperatures (above approximately 125 °C). Here we show that a cell made of a  $\text{CsHSO}_4$  electrolyte membrane (about 1.5 mm thick) operating at 150–160 °C in a  $\text{H}_2/\text{O}_2$  configuration exhibits promising electrochemical performances: open circuit voltages of 1.11 V and current densities of 44 mA cm<sup>-2</sup> at short circuit. Moreover, the solid-acid properties were not affected by exposure to humid atmospheres. Although these initial results show promise for applications, the use of solid acids in fuel cells will require the development of fabrication techniques to reduce electrolyte thickness, and an assessment of possible sulphur reduction following prolonged exposure to hydrogen.

Solid acids are compounds such as  $\text{KHSO}_4$  whose chemistry and properties lie between those of a normal acid (such as  $\text{H}_2\text{SO}_4$ ) and a normal salt (such as  $\text{K}_2\text{SO}_4$ )<sup>1</sup>. They are typically comprised of oxyanions—for example,  $\text{SO}_4$  or  $\text{SeO}_4$ —that are linked together via O–H···O hydrogen bonds. Within this category the  $\text{MHXO}_4$  and  $\text{M}_3\text{H}(\text{XO}_4)_2$  compounds, where M = Cs, NH<sub>4</sub>, Rb, and X = S or Se, are known to undergo a “superprotonic” phase transition<sup>2–9</sup>. On passing through the transition the conductivity jumps by several orders of magnitude to a value of 10<sup>-3</sup> to 10<sup>-2</sup> Ω<sup>-1</sup> cm<sup>-1</sup>, and the

materials often become quite ductile. The transition temperatures typically lie between 50 and 150 °C, and the activation energy for transport in the high-temperature phase is 0.3 to 0.45 eV (Fig. 1). A key feature of the proton transport process is that, unlike polymeric systems, it does not require humid atmospheres.

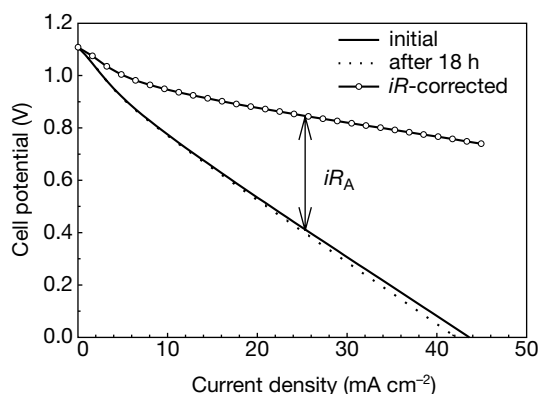
Although superprotonic solid acids have excellent proton transport properties, they have not been considered for applications, not only because of their water solubility and unusual mechanical properties, but also because they simply have not come to the attention of the fuel-cell community. In this work, we have selected the solid acid  $\text{CsHSO}_4$  for 'proof of principle' investigations in fuel-cell applications. This particular acid has been chosen because it is relatively well characterized and its superprotonic phase has a reasonably wide temperature range of stability, from the transition at 141 °C to decomposition/melting at approximately 200–230 °C (refs 5, 6). Membrane electrode assemblies (MEAs) were prepared as follows. A layer of the solid acid, synthesized from an aqueous solution of  $\text{Cs}_2\text{CO}_3$  and  $\text{H}_2\text{SO}_4$ , was sandwiched between two electrocatalysis layers comprised of  $\text{CsHSO}_4$ , Pt black, carbon black and a volatile organic in a mass ratio of 6:10:1:1 (approximate volume ratio of 4:1:2:2). These layers were, in turn, placed between two sheets of porous, graphite, current collectors. The entire assembly was uniaxially pressed at 490 MPa, to yield a dense electrolyte membrane (1–1.5 mm in thickness) with good mechanical contact to the electrocatalyst layers. Fuel-cell polarization curves were collected at slightly raised temperatures from MEAs placed in a standard graphite test station. Upon heating to the measurement temperature, the organic phase in the electrode evaporated, leaving behind a porous electrode structure.



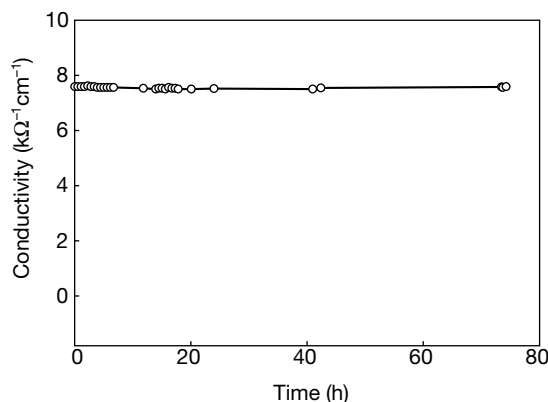
**Figure 1** The conductivities of selected solid acids, shown in Arrhenius form. Sources as indicated on the figure.

The current densities of such a single-cell fuel cell, exposed to  $\text{H}_2\text{O}$ -saturated  $\text{H}_2$  at the anode and  $\text{H}_2\text{O}$ -saturated  $\text{O}_2$  at the cathode and held at 160 °C, are presented in Fig. 2. The platinum content or 'loading' was  $18 \text{ mg cm}^{-2}$  and the membrane thickness 1.37 mm. The solid curve reflects the initial measurement and the dotted curve shows the measurement after 18 h of exposure to humid air. The anode and cathode gases were humidified during fuel-cell operation only in order to establish explicitly the open-circuit potential. Similar results were obtained for operation with dry gases. It is immediately apparent that despite the water solubility of  $\text{CsHSO}_4$  and its ability to undergo extensive plastic deformation at high temperatures, stable fuel-cell performance is possible. The slight drop in performance with time is probably due to degradation at the anode as a result of interactions with  $\text{H}_2$ , as described below, and not a result of interactions with  $\text{H}_2\text{O}$  or any lack of dimensional integrity.

The open-circuit potential of the cell of Fig. 2 is 1.11 V. This compares favourably with the theoretical value of 1.22 V, expected for the conditions  $T = 160^\circ\text{C}$ , and the partial pressures  $p_{\text{O}_2} \approx 1 \text{ atm}$ ,  $p_{\text{H}_2} \approx 1 \text{ atm}$ , and  $p_{\text{H}_2\text{O}} = 3.13 \times 10^{-2} \text{ atm}$  (ref. 10). The difference between the measured and theoretical open circuit voltages (OCV) is probably due to gas permeability through residual pores within the polycrystalline electrolyte, or to slow gas leaks through the experimental apparatus. Nevertheless, the OCV obtained here is significantly higher than that typically observed in polymer-based



**Figure 2** Cell voltage versus current density (polarization curve) for a  $\text{CsHSO}_4$  fuel cell. It was operated at 160 °C under moist  $\text{H}_2$  and moist  $\text{O}_2$  at the anode and cathode respectively (total pressure of 1 atm). Platinum loading of  $18 \text{ mg cm}^{-2}$ . Solid line, initial measurement; dotted line, measurement after 18 h of exposure to humidified air at both electrodes; dashed-dotted line,  $iR$ -corrected result for initial measurement.

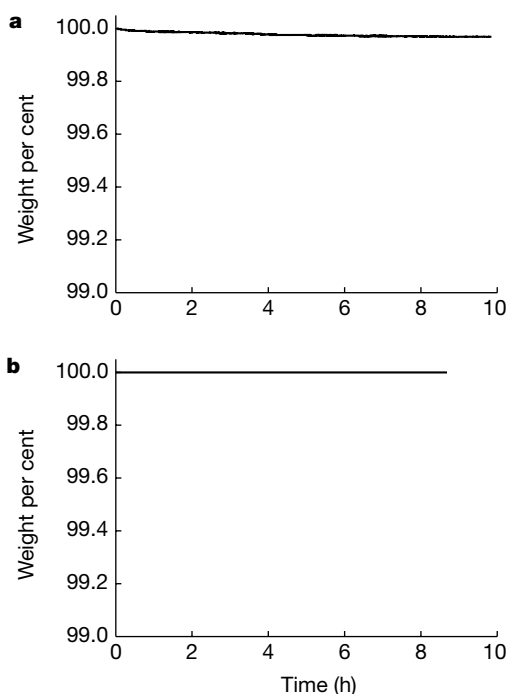


**Figure 3** Conductivity of  $\text{CsHSO}_4$  as a function of time at 146 °C and under humidified air.  $p_{\text{H}_2\text{O}} = 3.13 \times 10^{-2} \text{ atm}$ .



fuel cells: 0.9–1.0 V (ref. 11). The humidification requirements of the polymer, as well as its inherent permeability to O<sub>2</sub> and H<sub>2</sub> by molecular dissolution and transport, act to lower the maximum OCV achievable. In the case of the solid acid, we believe that improved processing will result in OCV values as high as the theoretical ones, as has been reported, for example, for KH<sub>2</sub>PO<sub>4</sub> (ref. 12) and routinely for solid-oxide fuel cells<sup>13</sup>. Fuel-cell experiments in which dry gases were used yielded slightly higher OCVs up to 1.25 V, as we expected.

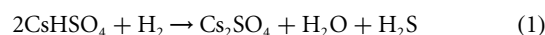
The drop in fuel-cell voltage with increasing current density (Fig. 2) has many causes<sup>11</sup>; the primary of these, for a relatively thick electrolyte as was used in this experiment, is the resistance of the membrane. This term is given by  $iR_A$ , where  $i$  is the current density, and  $R_A$  is the area specific resistance, or  $t/\sigma$ , where  $t$  is the electrolyte thickness and  $\sigma$  is the conductivity. Thus, one expects a linear drop in voltage as a function of current density, the magnitude of which depends only on the properties and geometry of the electrolyte. The characteristics of the superprotonic phase transition of CsHSO<sub>4</sub> are well-established, but the reported conductivity at 160 °C varies from  $3.70 \times 10^{-3} \Omega^{-1} \text{cm}^{-1}$  (ref. 6) to  $2.08 \times 10^{-3} \Omega^{-1} \text{cm}^{-1}$  (ref. 5). Measurements in our laboratory under H<sub>2</sub>O-saturated air indicate a conductivity of about  $8 \times 10^{-3} \Omega^{-1} \text{cm}^{-1}$ . Using this latter value and the electrolyte thickness of 1.37 mm, the expected slope in Fig. 2 is  $-17.1 \Omega \text{cm}^2$ , significantly smaller than the measured value of  $-22.7 \Omega \text{cm}^2$ . Additional resistance effects are probably due to slow charge transfer through the electrodes (also rather thick) and various contact resistances. The dashed-dotted curve in Fig. 2 ( $E_{\text{meas}} + iR_A$ ) shows the magnitude of the losses due to all non-electrolyte factors. The electrode resistance was fairly insensitive to the Pt loading over the range examined, 3.8 mg cm<sup>-2</sup> to 38 mg cm<sup>-2</sup> (see Supplementary Information), which suggests that the electrode performance is not limited by the catalyst content, but rather by the electrode microstructure. Although these preliminary results demonstrate that the lower limit has not been reached, improved fabrication techniques will be necessary to further reduce Pt loadings.



**Figure 4** Weight as a function of time of CsHSO<sub>4</sub> at 160 °C. **a**, Under flowing H<sub>2</sub>; **b**, under flowing O<sub>2</sub>.

In this work we used rather thick membranes to obtain impermeable MEAs and stable OCVs. In order to achieve an area-specific resistivity that is directly comparable to that of Nafion membranes (typically 0.025 to 0.0875  $\Omega \text{cm}^2$  at 90 °C and under highly humidified atmospheres for membranes 50–175  $\mu\text{m}$  in thickness<sup>11</sup>) a dense CsHSO<sub>4</sub> membrane with a thickness of the order of 2–20  $\mu\text{m}$  would be necessary. Such films may not be immediately achievable, but routine fabrication of solid-oxide electrolytes with comparable dimensions<sup>14</sup> renders this target quite realistic.

In order to assess the stability of CsHSO<sub>4</sub> under fuel-cell conditions, several longevity experiments were carried out. To examine the impact of humid conditions, its conductivity was measured at 146 °C under H<sub>2</sub>O-saturated air by impedance spectroscopy. For this experiment, electrodes were attached in a manner identical to that used for MEA fabrication. To examine the impact of reducing and oxidizing atmospheres, thermal gravimetric analysis was carried out under flowing H<sub>2</sub> and flowing O<sub>2</sub>, respectively. The results of the conductivity measurements are presented in Fig. 3, and the results of the thermal analyses are displayed in Fig. 4. The data in Fig. 3 demonstrate that the conductivity of CsHSO<sub>4</sub> is absolutely stable over tens of hours in humid atmospheres. Similarly, the results in Fig. 4 reveal that the solid acid is completely stable under oxidizing conditions, as expected. Under reducing conditions, however, the material has lost 0.03 wt% over 10 h. This weight loss is probably due to gradual sulphur reduction according to reaction (1), which, in turn, may be responsible for the slight loss in fuel-cell performance (Fig. 2) via the production of the catalyst poison, H<sub>2</sub>S.



The potential for sulphur reduction in solid-acid sulphates indicates that ultimately alternative oxyanion systems such as phosphates (for example, CsH<sub>2</sub>PO<sub>4</sub>; ref. 15) and complex heteropolyacids (for example, H<sub>3</sub>PW<sub>12</sub>O<sub>40</sub>·nH<sub>2</sub>O; ref. 16), which are not susceptible to reduction, may be the solid acid of choice for fuel-cell applications.

The results shown in Figs 2–4 demonstrate that inorganic, water-soluble solid acids can be used successfully in H<sub>2</sub>/O<sub>2</sub> fuel cells, and that MEAs fabricated from these materials yield higher OCVs than those obtained from polymer-electrolyte fuel cells. Higher OCVs, in turn, may lead to better overall system efficiencies. Further advantages that may ultimately result from such inorganic, non-hydrated electrolytes are: greater tolerance of the catalysts to CO due to slightly raised temperatures of operation; reduction in system complexity by elimination of costly temperature and humidity monitoring and control hardware; and applicability in direct (vapour) methanol fuel cells without methanol cross-over losses<sup>17</sup>. Several challenges remain before these advantages can be realized, including: (1) processing of thin, impermeable solid-acid membranes, (2) enhancement of electrode performance, and (3) system design to protect the electrolyte from liquid water during both intentional and inadvertent fuel-cell shut-off. Moreover, phosphate or other oxyanion-based superprotonic conductors may be preferable for long-term stability over those based on sulphates or selenates. □

Received 19 September 2000; accepted 22 February 2001.

1. Wells, A. F. *Structural Inorganic Chemistry* 3rd edn, 300–303 (Clarendon, Oxford, 1984).
2. Haile, S. M., Lentz, G., Kreuer, K.-D. & Maier, J. Superprotonic conductivity in Cs<sub>3</sub>(HSO<sub>4</sub>)<sub>2</sub>(H<sub>2</sub>PO<sub>4</sub>). *Solid State Ion.* **77**, 128–134 (1995).
3. Haile, S. M., Calkins, P. M. & Boysen, D. Superprotonic conductivity in  $\beta$ -Cs<sub>3</sub>(HSO<sub>4</sub>)<sub>2</sub>(H<sub>2</sub>(P(S)O<sub>4</sub>)). *Solid State Ion.* **97**, 145–151 (1997).
4. Chisholm, C. R. I. & Haile, S. M. Superprotonic behavior of CsHSO<sub>4</sub>–CsH<sub>2</sub>PO<sub>4</sub> system. *Solid State Ion.* **136–137**, 229–241 (2000).
5. Baranov, A. I., Shuvalov, L. A. & Shchagina, N. M. Superior conductivity and phase transitions in CsHSO<sub>4</sub> and CsHSeO<sub>4</sub> crystals. *JETP Lett.* **36**, 459–462 (1982).
6. Norby, T., Friesel, M. & Mellander, B. E. Proton and deuteron conductivity in CsHSO<sub>4</sub> and CsDSO<sub>4</sub> by in situ isotopic exchange. *Solid State Ion.* **77**, 105–110 (1995).
7. Pawlowski, A., Pawlaczyk, Cz. & Hilzner, B. Electric conductivity in crystal group Me<sub>3</sub>H(SeO<sub>4</sub>)<sub>2</sub> (Me: NH<sub>4</sub><sup>+</sup>, Rb<sup>+</sup>, Cs<sup>+</sup>). *Solid State Ion.* **44**, 17–19 (1990).

8. Ramasastry, C. & Ramaiah, A. S. Electrical conduction in  $\text{Na}_3\text{H}(\text{SO}_4)_2$  crystals. *J. Mater. Sci. Lett.* **16**, 2011–2016 (1981).
9. Chisholm, C. R. L., Merinov, B. V. & Haile, S. M. High temperature phase transitions in  $\text{K}_3\text{H}(\text{SO}_4)_2$ . *Solid State Ion.* (in the press).
10. Gaskell, D. R. *Introduction to Metallurgical Thermodynamics* 2nd edn, 574 & 586 (McGraw-Hill, Washington DC, 1981).
11. Srinivasan, S., Velev, O. A., Parthasarathy, A., Manko, D. J. & Appleby, A. J. High energy efficiency and high power density proton exchange membrane fuel cells—electrode kinetics and mass transport. *J. Power Sources* **36**, 299–320 (1991).
12. Croce, F. & Cigna, G. Determination of the protonic transference number for  $\text{KH}_2\text{PO}_4$  by electromotive force measurements. *Solid State Ion.* **6**, 201–202 (1982).
13. Minh, N. Q. Ceramic fuel cells. *J. Am. Ceram. Soc.* **76**, 563–588 (1993).
14. Minh, N. Q. & Horne, C. R. in *Proc. 14th Riso Int. Symp. on Materials Science: High Temperature Electrochemical Behaviour of Fast Ion and Mixed Conductors* (eds Poulsen, F. W., Bentzen, J. J., Jacobsen, T., Skou, E. & Østergård, M. J. L.) 337–341 (Riso National Laboratory, Roskilde, 1993).
15. Baranov, A. I., Khiznichenko, V. P., Sandler, V. A. & Shuvalov, L. A. Frequency dielectric dispersion in the ferroelectric and superionic phases of  $\text{CsH}_2\text{PO}_4$ . *Ferroelectrics* **81**, 183–186 (1988).
16. Slade, R. C. T. & Omana, M. J. Protonic conductivity of 12-tungstophosphoric acid (TPA,  $\text{H}_3\text{PW}_{12}\text{O}_{40}$ ) at elevated temperatures. *Solid State Ion.* **58**, 195–199 (1992).
17. Haile, S. M., Narayanan, S. R., Chisholm, C. & Boysen, D. Proton conducting membrane using a solid acid. US patent 09/439,377 (15 Nov. 2000).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

## Acknowledgements

We thank J. Snyder for assistance with computer automation of the fuel cell test station and S. R. Narayanan for technical discussions. Financial support has been provided by the California Institute of Technology through the Yuen Grubstake entrepreneurial fund.

Correspondence and requests for materials should be addressed to S.M.H. (e-mail: [smhaile@caltech.edu](mailto:smhaile@caltech.edu)).

# Self-assembly of mesoscopically ordered chromatic polydiacetylene/silica nanocomposites

Yunfeng Lu<sup>†‡§</sup>, Yi Yang<sup>†</sup>, Alan Sellinger<sup>‡§</sup>, Mengcheng Lu<sup>†</sup>, Jinman Huang<sup>†</sup>, Hongyou Fan<sup>‡</sup>, Raid Haddad<sup>†</sup>, Gabriel Lopez<sup>†</sup>, Alan R. Burns<sup>‡</sup>, Darryl Y. Sasaki<sup>‡</sup>, John Shelnutt<sup>‡</sup> & C. Jeffrey Brinker<sup>†‡§</sup>

<sup>†</sup> The University of New Mexico Center for Micro-Engineered Materials and Department of Chemical and Nuclear Engineering, Albuquerque, New Mexico 87131, USA

<sup>‡</sup> Sandia National Laboratories, Advanced Materials Laboratory, 1001 University Blvd SE, Albuquerque, New Mexico 87106, USA

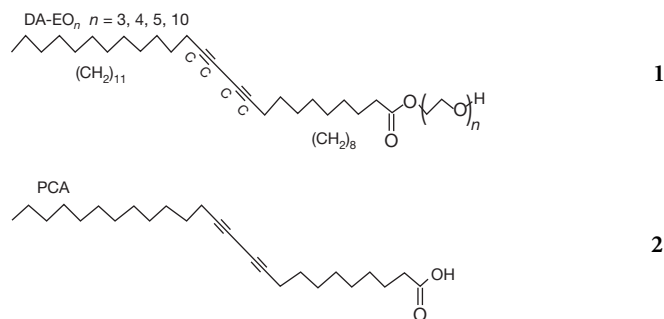
<sup>§</sup> These authors contributed equally to this work.

Nature abounds with intricate composite architectures composed of hard and soft materials synergistically intertwined to provide both useful functionality and mechanical integrity. Recent synthetic efforts to mimic such natural designs have focused on nanocomposites<sup>1–5</sup>, prepared mainly by slow procedures like monomer or polymer infiltration of inorganic nanostructures<sup>6,7</sup> or sequential deposition<sup>8,9</sup>. Here we report the self-assembly of conjugated polymer/silica nanocomposite films with hexagonal, cubic or lamellar mesoscopic order using polymerizable amphiphilic diacetylene molecules as both structure-directing agents and monomers. The self-assembly procedure is rapid and incorporates the organic monomers uniformly within a highly

ordered, inorganic environment. Polymerization results in polydiacetylene/silica nanocomposites that are optically transparent and mechanically robust. Compared to ordered diacetylene-containing films prepared as Langmuir monolayers<sup>10</sup> or by Langmuir–Blodgett deposition<sup>10</sup>, the nanostructured inorganic host alters the diacetylene polymerization behaviour, and the resulting nanocomposite exhibits unusual chromatic changes in response to thermal, mechanical and chemical stimuli. The inorganic framework serves to protect, stabilize, and orient the polymer, and to mediate its function. The nanocomposite architecture also provides sufficient mechanical integrity to enable integration into devices and microsystems.

Owing to extended  $\pi$ -electron delocalization along their backbones, conjugated organic polymers exhibit electronic and optical properties of interest for applications ranging from light-emitting diodes to biomolecular sensors<sup>11</sup>. For example, in blue-coloured polydiacetylene, the optical absorption blue-shifts dramatically when stress is applied to the backbone through the pendant side chains, and this thermally, mechanically, or chemically induced chromatic (blue  $\rightarrow$  red) response has been explored as a colorimetric transduction scheme in a variety of chemically and physically based sensor designs<sup>12,13</sup>. Further improvements of the electronic and optical performance of conjugated polymer devices may require polymer incorporation in nano-engineered architectures<sup>14</sup> that could provide alignment, control charge and energy transfer, mediate conformational changes and prevent oxidation. Recently control of energy transfer was demonstrated in a poly[2-methoxy-5-(2'-ethyl-hexyloxy)-1,4-phenylene vinylene] (MEH-PPV)/silica nanocomposite<sup>7</sup>. However, this nanocomposite, prepared by MEH-PPV infiltration of a pre-formed, oriented, hexagonal, silica mesophase, was heterogeneous, exhibiting two distinct conjugated polymer environments, that is, polymers inside and outside the hexagonally arranged pore channels of the silica particles. In general, because polymer infiltration into a pre-formed porous nanostructure depends on the partitioning of the polymer from the solvent, we expect it to be difficult to control polymer concentration, orientation and uniformity in the corresponding nanocomposite. Further, when the nanostructure pore size is less than the radius of gyration of the solvated polymer, infiltration proceeds by a worm-like motion referred to as reptation, requiring long processing times at elevated temperatures<sup>7</sup>.

We use a series of oligoethylene glycol functionalized diacetylenic (DA-EO<sub>n</sub>) surfactants (structure 1, with  $n = 3, 4, 5$ , or 10), prepared by coupling ethylene glycols with the acid chloride of PCA (structure 2)



both as amphiphiles to direct the self-assembly of thin film silica mesophases<sup>15</sup> and as monomeric precursors of the conjugated polymer, polydiacetylene (PDA). Beginning with a homogeneous solution of silicic acid and surfactant prepared in a tetrahydrofuran (THF)/water solvent with initial surfactant concentration  $c_0$  much less than the critical surfactant micelle concentration CMC, we use evaporative dip-coating, spin-coating, or casting procedures to prepare thin films on silicon (100) or fused silica substrates (Fig. 1). During deposition, preferential evaporation of THF

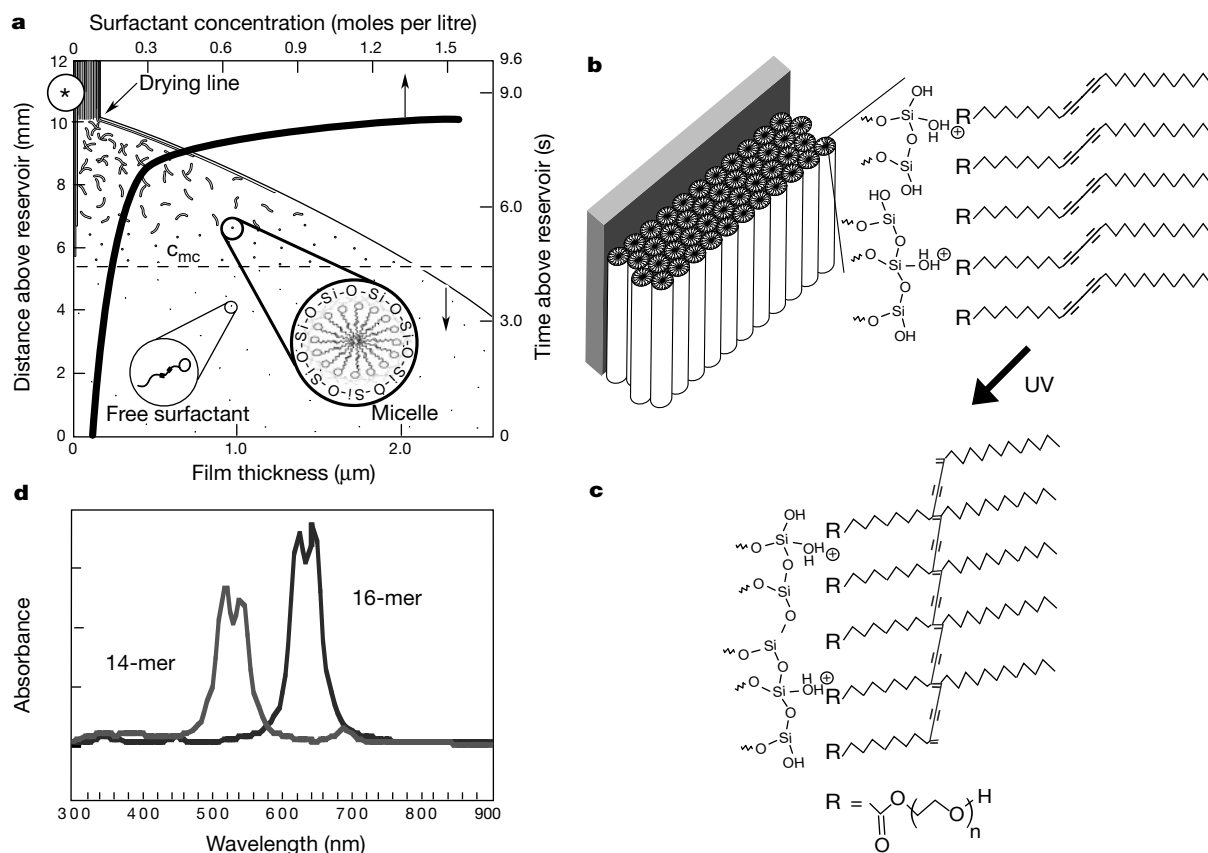
<sup>§</sup> Present address: Tulane University Chemical Engineering Department, New Orleans, Louisiana 70118, USA (Y.L.); Canon Research and Development Center Americas, San Jose, California 95134, USA (A.S.).

concentrates the depositing film in water and nonvolatile silica and surfactant species. The progressively increasing surfactant concentration drives self-assembly of diacetylene/silica surfactant micelles and their further organization into ordered, three-dimensional, liquid crystalline mesophases. Ultraviolet-light-initiated polymerization of the DA units, accompanied by catalyst-promoted siloxane condensation, topochemically converts the colourless mesophase into the blue PDA/silica nanocomposite, preserving the highly ordered, self-assembled architecture.

The choice of surfactant greatly influences the resultant mesostructure. This is evident from the X-ray diffraction (XRD) patterns and TEM micrographs shown in Figs 2 and 3 for nanocomposites prepared from diacetylenic surfactants with tri ( $n = 3$ ), penta ( $n = 5$ ) and decaethylene ( $n = 10$ ) glycol head groups. Increasing values of  $n$  increase the surfactant head group area  $a_0$ . This in turn reduces the value of the surfactant packing parameter,  $g = v/a_0l$ , where  $v$  is the surfactant volume and  $l$  the tail length<sup>16</sup>, favouring the formation of progressively higher-curvature mesophases: lamellar ( $n = 3$ )  $\rightarrow$  hexagonal ( $n = 5$ )  $\rightarrow$  cubic ( $n = 10$ ). From the highly ordered nanocomposite mesostructures observed by transmission electron microscopy (TEM), we infer that the surfactant monomers/structure-directing agents are uniformly

organized into precise spatial arrangements before polymerization. These arrangements establish the proximity (topochemistry) of the reactive diacetylenic moieties and thus strongly influence the PDA polymerization process. This is best illustrated by comparing their polymerization behaviour, as shown by the blue (or red) colour of nanocomposite films prepared with different mesostructures (that is, the hexagonal, cubic, and lamellar mesostructures shown in Fig. 3) and contrasting these behaviours with those of planar self-assembled monolayer and trilayer films formed by Langmuir–Blodgett deposition of the neat DA surfactants.

Figure 4 shows a patterned blue PDA/silica nanocomposite film, with a hexagonal mesostructure (prepared using structure 1, with  $n = 5$ ), formed by ultraviolet exposure through a mask; the corresponding patterned red film was formed subsequently by heating to 100 °C (Fig. 4b). Whereas lamellar mesophases (prepared using structure 1 with  $n = 3$ ) show qualitatively similar behaviour, cubic mesostructures (prepared using structure 1 with  $n = 10$ ) and Langmuir monolayers and trilayers (prepared using neat 1 with  $n = 3, 5$  or 10) remain colourless upon ultraviolet exposure and during heating. The different behaviour of lamellar and Langmuir films emphasize the importance of the nanostructured inorganic host on PDA polymerization. In both systems the diacetylenic



**Figure 1** Schematic representation of the PDA/silica nanocomposite evaporation-induced self-assembly process. **a**, Steady-state film thickness and surfactant concentration profiles developed during dip-coating, with vertical axes representing distance and time above reservoir surface and horizontal axes showing film thickness and surfactant (structures 1 or 2) concentration for  $c_0 = 0.034$  M. Beginning with a homogeneous solution prepared with  $c_0 < \text{CMC}$ , preferential THF evaporation induces micellization and further self-organization of silica–surfactant composite micelles into ordered thin-film mesophases. The shape and concentration of the DA surfactants influence the mesophase established at the drying line (lamellar, hexagonal or cubic). **b**, Section near asterisk shows a hexagonal mesostructure and hypothetical arrangement of DA

surfactants adjacent to the cylindrically structured silicic acid framework. **c**, Hypothetical structure of polymerized PDA/silica nanocomposite formed upon exposure to ultraviolet light and continued acid-catalysed siloxane condensation. **d**, Ultraviolet–visible spectra of cyclic polydiacetylene oligomers determined by INDO/S (Intermediate Neglect Differential Orbital Spectra) quantum calculations after energy-optimization and 10 ps of molecular dynamics. The 16-membered oligomer with diameter of about 2.4 nm absorbs in the red region of the visual spectrum with spectral characteristics commensurate with the blue form of PDA (compare with Fig. 4e). The higher-curvature 14-membered PDA oligomer is strained and has a blue-shifted absorption spectrum consistent with the red form of PDA (compare with Fig. 4e).

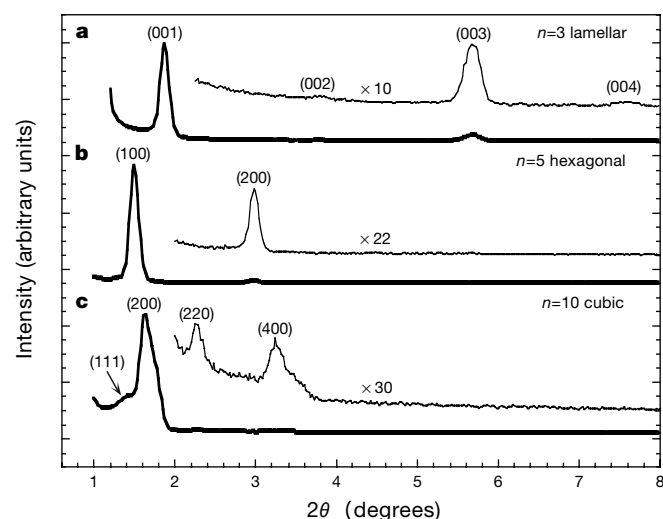


surfactants are organized into highly oriented planar configurations with the EO headgroups disposed toward the hydrophilic interface, either water (Langmuir films) or polysilicic acid (nanocomposites). Despite these similar organizations, we suggest that Langmuir films do not polymerize because the reactive DA moieties are spaced too far apart, as indicated by the molecular areas measured at 30 mN m<sup>-1</sup> in a Langmuir trough: DA-EO<sub>1</sub> = 24 Å<sup>2</sup>, DA-EO<sub>3</sub> = 38 Å<sup>2</sup>, DA-EO<sub>5</sub> = 46 Å<sup>2</sup>. Closer spacing of the EO headgroups (and correspondingly the diacetylenic moieties) within the self-assembled nanocomposites (lamellar, hexagonal or cubic) is anticipated from the requirement for charge density matching at the R-EO<sub>n-y</sub>[EO·H<sub>3</sub>O<sup>+</sup>]<sub>y</sub>·yX<sup>-</sup>·wI<sup>δ+</sup> interface (where R = alkyl chain,  $y \leq n$ , X = Cl<sup>-</sup>, I<sup>δ+</sup> = the silica framework carrying a partial charge of δ<sup>+</sup>, and w is the 'concentration' of framework needed for charge balance) that reduces the optimal EO headgroup area *a*<sub>0</sub> (ref. 17). This closer spacing is reflected in measurements of the biaxial stress resulting from PDA polymerization, where, for nanocomposites, we observe development of compressive stress upon polymerization (film turns blue) and during the solvatochromic or thermochromic blue → red transformation; this indicates a net expansion of the PDA relative to the silica host (see Fig. A in the Supplementary Information). The corresponding neat DA surfactants prepared as Langmuir monolayers do not polymerize and develop no stress upon ultraviolet irradiation. Polymerizable DA Langmuir monolayers and multilayers (prepared using surfactants with smaller headgroups) contract upon polymerization and the magnitude of this

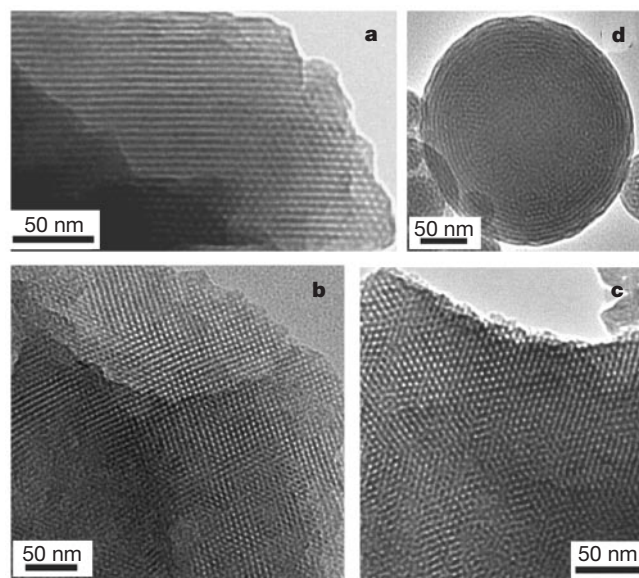
contraction is used to assess the extent of polymerization<sup>18</sup>.

The importance of the proximity of the DA moieties on polymerization is further illustrated by our inability to polymerize DA within the cubic mesophase prepared with *n* = 10 (see Fig. 3b and c). The large EO<sub>10</sub> surfactant headgroups are necessary to direct the formation of the cubic mesophase, but they also serve as spacers preventing polymerization. Introduction of structure 1, with *n* = 3 or 5 or structure 2 as co-surfactants with smaller headgroups (30–50 wt% relative to EO-DA<sub>10</sub>) allows us to form cubic mesophases and to polymerize the mixed surfactant assemblies into the blue and red forms of PDA. (We expect that, owing to phase separation<sup>19</sup>, this is not possible in Langmuir monolayers or Langmuir–Blodgett films).

The polymerization of the surfactant is highly dependent upon the topological alignment of diacetylenic units within the supramolecular assembly<sup>20,21</sup>. Additionally, for colorimetric materials to form (especially blue-coloured materials), the degree of polymerization and conjugation length of the 'ene'–'yne' backbone must be considerable<sup>22</sup>. The micellar structures that template the silica sol-gel material must thus contain highly oriented, densely packed surfactants. This packing will allow facile topochemical polymerization of the diacetylene units to give coloured polydiacetylene.



**Figure 2** X-ray diffraction (XRD) patterns of nanocomposite thin films prepared using DA-EO<sub>*n*</sub> (where *n* = 3, 5, and 10) surfactants. Increasing values of *n* increase the surfactant head group favouring the formation of progressively higher-curvature mesophases: lamellar (*n* = 3) → hexagonal (*n* = 5) → cubic (*n* = 10). **a**, Film prepared using 1.86% of DA-EO<sub>3</sub> (wt% DA-EO<sub>3</sub> relative to the solution weight without surfactant) shows a lamellar mesostructure with (001) diffraction peak at 48 Å. The presence of higher-order diffraction peaks (002), (003) and (004) indicate the formation of highly ordered alternating silica/PDA layers. **b**, Film prepared with DA-EO<sub>5</sub> (2.23%) shows a hexagonal mesophase with strong (100) and (200) diffraction peaks at 56 Å and 28 Å, respectively. **c**, Film prepared using 2.23% DA-EO<sub>10</sub> surfactant exhibits a cubic mesophase with unit-cell parameter *a* = 108.1 Å. The presence of the (111) diffraction peak at 62.4 Å and the (200) diffraction peak at 54.1 Å clearly establishes a face-centred or primary cubic structure. The presence of higher-order peaks (220) at 38.2 Å and (400) at 27.3 Å provide further evidence of the primary or face-centred cubic mesophase. Introduction of structure 1, with *n* = 3 or 5 or structure 2, as co-surfactants with smaller headgroups (30–50 wt% relative to EO-DA<sub>10</sub>) was necessary to polymerize PDA in the blue or red forms within the cubic mesophase. The XRD patterns were essentially unchanged by these co-surfactant additions.



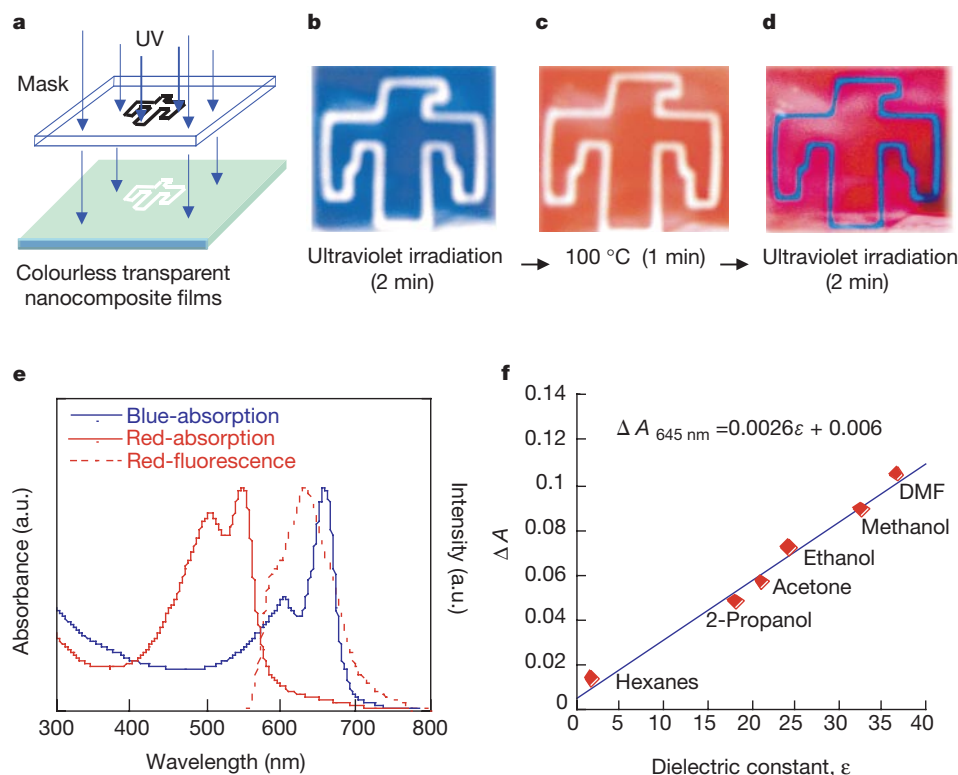
**Figure 3** Representative transmission electron microscope (TEM) images of nanocomposite thin films (prepared as in Fig. 2b and c) and particles (formed by a related aerosol-assisted evaporation-induced self-assembly approach<sup>32</sup>). **a**, Hexagonally ordered nanocomposite prepared using 2.23% DA-EO<sub>5</sub> surfactant, showing a striped [110]-oriented mesophase with repeat distance around 48 Å and a region of the corresponding hexagonally patterned mesostructure characteristic of the [001]-orientation (right edge). The smaller spacing measured by TEM versus XRD (48 Å versus 56 Å, compare **a** and Fig. 2b) may be due to the over-focus condition used to achieve high contrast in TEM or electron-beam-induced shrinkage (normal to the substrate) which can occur in general for uncalcined, mesostructured materials. **b, c**, Two orientations of a cubic nanocomposite film prepared using 2.23% DA-EO<sub>10</sub> surfactant. A highly ordered [100] orientation with unit-cell parameter at 108 Å, agreeing well with the XRD results. **c**, The corresponding [111] orientation with repeat distance of 62 Å. As for the corresponding XRD patterns, the cubic mesostructures revealed by TEM were quite comparable with and without the addition of co-surfactants (structure 1, with *n* = 3 or 5, or structure 2) needed to polymerize the blue or red forms of PDA within the cubic mesostructure. **d**, PDA/silica nanocomposite particles prepared using an aerosol-assisted evaporation-induced self-assembly technique<sup>32</sup>. The particle exhibits an ordered multi-lamellar exterior and a disordered worm-like mesostructured interior, consistent with a radially directed self-assembly process.

Figure 1 illustrates a postulated length-wise polymerization of surfactants within a cylindrical rod-like micelle to yield the blue form of PDA. Circumferential polymerization, although favoured on the basis of proximity of reactive DA moieties, would impose a high curvature on the PDA backbone because of the small pore size of the mesostructured host (2.5–3.0 nm in diameter, see Fig. B in the Supplementary Information). Although DA surfactants organized into ellipsoidal liposomes (40 nm × 15 nm) are known to polymerize in the blue and red forms<sup>23</sup> we questioned whether circumferential arrangements of DA surfactants (within spherical or cylindrical micelles) could polymerize with the proper ‘ene’–‘yne’ conjugation to produce coloured materials. Molecular mechanics and dynamics calculations indicate that cyclic arrangements of up to 19 DA monomers around the periphery of the 3.0-nm-diameter pore are possible and that an energy-optimized 16-membered cyclic PDA oligomer can exhibit absorption behaviour consistent with the blue form of PDA (see Fig. 1d). Helical conformations<sup>24</sup> represent an alternative configuration that might satisfy DA proximity requirements without severely disrupting the conjugation. In general, nonlinear conformations may be needed for polymer propagation within these self-assembled nano-structured hosts. Such conformations are presumably rare or impossible for Langmuir monolayers and Langmuir–Blodgett films.

The PDA/silica nanocomposite films are optically transparent and mechanically robust. Modulus values measured by nano-indentation ( $3.50 \pm 1.00$  GPa) are compared to those of calcined mesoporous silica films that have been successfully integrated into microelectronic devices as low dielectric constant films (Y.L. and

C.J.B., unpublished work). Nitrogen sorption experiments performed using a surface acoustic wave technique<sup>25</sup> show the nanocomposite films to be nonporous to nitrogen at  $-196^\circ\text{C}$  (Type II sorption isotherm, see Fig. B of the Supplementary Information), meaning that the PDA completely fills the pore channels and that the composite architecture may impart some degree of oxidation resistance (important for extension to other conjugated polymer nanocomposites).

The blue PDA/silica nanocomposites thus exhibit solvatochromic, thermochromic, and mechanochromic properties. When contacted with the series of polar solvents—2-propanol, acetone, ethanol, methanol and dimethylformamide—the films transform to the red, fluorescent form and the differential absorbance ( $A$ ) at 645 nm ( $A_{\text{blue}} - A_{\text{red}}$ ) scales linearly with the dielectric constant of the solvent:  $A_{645\text{ nm}} = 0.003$  (dielectric constant)  $- 0.005$ ; for a correlation coefficient of 0.992 (see Fig. 4f), suggesting applications in sensing. We suggest that polar solvents diffuse within the polar, hydrophilic  $\text{EO}_n$  pendant side chains. The accompanying solvation stresses are transferred to the PDA backbone (evidence for solvation stresses are presented in Fig. A of the Supplementary Information), reducing its conjugation length, and therefore inducing the blue  $\rightarrow$  red transformation<sup>26</sup> as understood by recent molecular mechanics simulations<sup>27</sup>. The solvatochromic behaviour shows very slow reversibility. Heating to temperatures in excess of  $47^\circ\text{C}$  can also cause the blue  $\rightarrow$  red transformation, and preliminary results have shown this thermochromic behaviour to be rapidly reversible (several seconds). This may arise from hydrogen-bonding interactions<sup>26</sup> of the pendant side chains with the silanol moieties



**Figure 4** Patterned polymerization induced by ultraviolet irradiation and the thermochromic and solvatochromic transition of a hexagonal PDA/silica nanocomposite film. **a**, Schematic representation of the ultraviolet patterning procedure. **b**, Patterned blue/colourless film formed by ultraviolet irradiation for 2 min. **c**, Patterned red/colourless film formed by heating the film in **b** to  $100^\circ\text{C}$  for 1 min. **d**, Patterned red/blue film formed by ultraviolet exposure of film in **c** for 2 min. **e**, Ultraviolet–visible spectra of the blue PDA/silica hexagonally ordered nanocomposite, the corresponding red film formed by heating at  $100^\circ\text{C}$ , and the fluorescence emission spectrum of the red film (excited at 500 nm).

Film formed from structure **1**, with  $n = 5$ . **f**, Differential absorbance at 645 nm resulting from the solvatochromic transition of blue PDA/silica nanocomposite films to the corresponding red films upon immersion of the blue films in the series of solvents: hexane, 2-propanol, acetone, ethanol, methanol, or dimethylformamide. Immersion times were 3 s followed by drying in nitrogen at room temperature. Without consideration of the non-polar solvent, hexane, the regression equation is  $A_{645\text{ nm}} = 0.003$  (dielectric constant)  $- 0.005$ , correlation coefficient = 0.992.

of the surrounding inorganic mesophase, supplying a restoring force and enabling recovery of the original side-chain orientation. Mechanical abrasion of the blue nanocomposite film causes local transformation to the red fluorescent form. Thus we envision mechanochromic barrier coatings that could sense excessive mechanical damage by changing from the blue form to the red fluorescent form.

The use of polymerizable surfactants as both structure-directing agents and monomers in the various evaporation-driven self-assembly schemes developed recently<sup>28,29</sup> represents a general, efficient route to the formation of robust and functional nanocomposites. Synthesis of surfactants with polymerizable thiophene, acetylene or phenylenevinylene groups should enable the self-assembly of conductive, conjugated polymer/silica nanocomposites in thin-film forms suitable for integration into devices. Conductive polymers confined within rod-like micellar channels of an electrical insulator conjure ideas of molecular wires. Unlike conduction measurements with two-dimensional films, such as bilayers<sup>30</sup> and monolayers<sup>31</sup>, hexagonally ordered nanocomposite mesophases are nearly one-dimensional structures that could allow efficient conductivity along the channel direction. □

## Methods

Precursor solutions were synthesized from tetraethylorthosilicate (TEOS, Si(OC<sub>2</sub>H<sub>5</sub>)<sub>4</sub>), diacetylenic surfactants (structure 1 with  $n = 3, 4, 5, 10$ ; structure 2, or combinations thereof) and HCl catalyst prepared in a THF/water solvent. The final reactant mole ratios were 1 TEOS:31.4–57.9 THF:4.96 H<sub>2</sub>O:0.013 HCl:0.06–1.41 DA surfactant. Films were prepared by casting, spin-coating at 2,000 r.p.m., or dip-coating at a rate of 40 cm min<sup>-1</sup>. Polymerization of PDA to the blue form was done by ultraviolet exposure at 266 nm for times ranging from 30 s to 30 min. Subsequent transformation to the red form was accomplished by heating at 100 °C for times ranging from 30 s to 2 min or by exposure to a solvent.

Molecular simulations were performed using Cerius<sup>2</sup> and Polygraf software and the Burchart–Dreiding 2.21 force field. Ultraviolet–visible spectra of the PDA oligomers were calculated from the transition wavelengths and oscillator strengths obtained from INDO/S calculations on the terminated polymer backbone structure obtained from the molecular dynamics calculation.

Received 3 November 2000; accepted 6 February 2001.

- Dagani, R. Putting the “nano” into composites. *Chem. Eng. News* **77**, 25–37 (1999).
- Giannelis, E. Polymer layered silicate nanocomposites. *Adv. Mater.* **8**, 29–35 (1996).
- Asefa, T., Yoshina-Ishii, C., MacLachlan, M. J. & Ozin, G. A. New nanocomposites: putting organic function “inside” the channel walls of periodic mesoporous silica. *J. Mater. Chem.* **10**, 1751–1755 (2000).
- Smith, R. C., Fischer, W. M. & Gin, D. L. Ordered poly(*p*-phenylenevinylene) matrix nanocomposites via lyotropic liquid-crystalline monomers. *J. Am. Chem. Soc.* **119**, 4092–4093 (1997).
- Sellinger, A. et al. Continuous self-assembly of organic-inorganic nanocomposite coatings that mimic nacre. *Nature* **394**, 256–260 (1998).
- Moller, K., Bein, T. & Fischer, R. X. Entrapment of PMMA polymer strands in micro- and mesoporous materials. *Chem. Mater.* **10**, 1841–1852 (1998).
- Nguyen, T.-Q., Wu, J., Doan, V., Schwartz, B. J. & Tolbert, S. H. Control of energy transfer in oriented conjugated polymer-mesoporous silica composites. *Science* **288**, 652–656 (2000).
- Kleinfeld, E. R. & Ferguson, G. S. Stepwise formation of multilayered nanostructural films from macromolecular precursors. *Science* **265**, 370–373 (1994).
- Keller, S. W., Kim, H.-N. & Mallouk, T. E. Layer-by-layer assembly of intercalation compounds and heterostructures on surfaces: Toward molecular “beaker” epitaxy. *J. Am. Chem. Soc.* **116**, 8817–8818 (1994).
- Sasaki, D. Y., Carpick, R. W. & Burns, A. R. High molecular orientation in mono- and trilayer polydiacetylene films imaged by atomic force microscopy. *J. Colloid Interf. Sci.* **229**, 490–496 (2000).
- Charych, D., Nagy, J., Spevak, W. & Bednarski, M. Direct colorimetric detection of a receptor-ligand interaction by a polymerized bilayer assembly. *Science* **261**, 585–588 (1993).
- Cheng, Q. & Stevens, R. Charge-induced chromatic transition of amino acid-derivatized polydiacetylene liposomes. *Langmuir* **14**, 1974–1976 (1998).
- Cheng, Q., Yamamoto, M. & Stevens, R. Amino acid terminated polydiacetylene lipid microstructures: morphology and chromatic transition. *Langmuir* **16**, 5333–5342 (2000).
- Mazumdar, S. Prospects for the polymer nanoengineer. *Science* **288**, 630–631 (2000).
- Lu, Y. F. et al. Continuous formation of supported cubic and hexagonal mesoporous films by sol-gel dip coating. *Nature* **389**, 364–368 (1997).
- Israelachvili, J. N. *Intermolecular and Surface Forces* Ch. 17 (Academic, San Diego, 1992).
- Monnier, A. et al. Cooperative formation of inorganic-organic interfaces in the synthesis of silicate mesostructures. *Science* **261**, 1299–1303 (1993).
- Day, D. & Ringsdorf, H. Polymerization of diacetylene carbonic acid monolayers at the gas-water interface. *J. Polym. Sci. Polym. Lett. Edn* **16**, 205–210 (1978).
- Gaines, G. L. (ed.) in *Insoluble Monolayers at Liquid-Gas Interfaces* 281–300 (John Wiley, New York, 1966).
- Menzel, H., Mowery, M. D., Cai, M. & Evans, C. E. Vertical positioning of internal molecular scaffolding within a single molecular layer. *J. Phys. Chem. B* **102**, 9550–9556 (1998).

- Collins, M. Optical-properties of poly-diacetylene monolayers. *J. Polym. Sci. B* **26**, 367–372 (1988).
- Kuriyama, K., Kikuchi, H. & Kajiyama, T. Molecular packings-photopolymerization behavior relationship of diacetylene Langmuir-Blodgett films. *Langmuir* **12**, 6468–6472 (1996).
- Spevak, W. et al. Polymerized liposomes containing C-glycosides of sialic acid: Potent inhibitors of influenza virus in vitro infectivity. *J. Am. Chem. Soc.* **115**, 1146–1147 (1993).
- Frankel, D. A. & O'Brien, D. F. Supramolecular assemblies of diacetylenic aldonamides. *J. Am. Chem. Soc.* **113**, 7436–7437 (1991).
- Frye, G. C., Ricco, A. J., Martin, S. J. & Brinker, C. J. in *Mater. Res. Soc.* (eds Brinker, C. J., Clark, D. E. & Ulrich, D. R.) 349–354 (Reno, Nevada, 1988).
- Patel, G. N., Chance, R. R. & Witt, J. D. A planar-non-planar conformational transition in conjugated polymer solutions. *J. Chem. Phys.* **70**, 4387–4392 (1979).
- Burns, A. R., Carpick, R. W., Sasaki, D. Y. & Shelnutt, J. A. Mechanochromism, shear force anisotropy, and molecular mechanics in polydiacetylene monolayers. *Tribol. Lett.* (in the press).
- Fan, H. et al. Rapid prototyping of patterned functional nanostructures. *Nature* **405**, 56–60 (2000).
- Brinker, C. J., Lu, Y., Sellinger, A. & Fan, H. Evaporation-induced self-assembly: nanostructures made easy. *Adv. Mater.* **11**, 579–585 (1999).
- Day, D. R. & Lando, J. B. Conduction in polydiacetylene bilayers. *J. Appl. Polym. Sci.* **26**, 1605–1612 (1981).
- Suzuoki, Y., Kimura, A. & Mizutani, T. in *Proc. 7th Int. Symp. Electrets* 850–855 (IEEE, New York, 1991).
- Lu, Y. et al. Aerosol-assisted self-assembly of spherical, silica nanoparticles exhibiting hexagonal, cubic and vesicular mesophases. *Nature* **398**, 223–226 (1999).

Supplementary information is available from Nature's World-Wide Website (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

## Acknowledgements

We thank J. Curro and F. van Swol for technical discussions and T. Buechert for the nano-indentation results. This work was supported by the US Department of Energy Basic Energy Sciences Program, the Sandia National Laboratories Laboratory-Directed Research and Development Program, the National Aeronautics and Space Administration, the UNM/NSF Center for Micro-engineered Materials, and the Defense Advanced Research Projects Agency Bio-Weapons Defense Program. TEM investigations were performed in the Department of Earth and Planetary Sciences at the University of New Mexico. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed-Martin Company, for the US DOE.

Correspondence and requests for materials should be addressed to C.J.B. (e-mail: [cjbrink@sandia.gov](mailto:cjbrink@sandia.gov)).

# Early Oligocene initiation of North Atlantic Deep Water formation

Richard Davies\*, Joseph Cartwright†, Jennifer Pike† & Charles Line\*

\* ExxonMobil International Limited, St Catherine's House, 2 Kingsway, PO Box 393, London WC2B 6WF, UK

† Department of Earth Sciences, PO Box 914, Cardiff University, Cardiff CF10 3YE, UK

Dating the onset of deep-water flow between the Arctic and North Atlantic oceans is critical for modelling climate change in the Northern Hemisphere<sup>1,2</sup> and for explaining changes in global ocean circulation throughout the Cenozoic era<sup>3</sup> (from about 65 million years ago to the present). In the early Cenozoic era, exchange between these two ocean basins was inhibited by the Greenland–Scotland ridge<sup>3,4</sup>, but a gateway through the Faeroe–Shetland basin has been hypothesized<sup>3,5</sup>. Previous estimates of the date marking the onset of deep-water circulation through this basin—on the basis of circumstantial evidence from neighbouring basins—have been contradictory<sup>5–9</sup>, ranging from about 35 to 15 million years ago. Here we describe the newly discovered Southeast Faeroes drift, which extends for 120 km parallel to the basin axis. The onset of deposition in this drift has been dated to the early Oligocene epoch (~35 million years ago) from a petroleum exploration borehole. We show that the drift was deposited under a southerly flow regime, and conclude that the



## initiation of deep-water circulation from the Norwegian Sea into the North Atlantic Ocean took place much earlier than is currently assumed in most numerical models of ancient ocean circulation.

The flux of cold, dense bottom water from the Arctic Ocean and Norwegian Greenland Sea exerts a fundamental influence on the world climate through its impact on deep-ocean circulation<sup>10,11</sup>. The present-day routes for this North Atlantic Deep Water (NADW), leaving the Norwegian–Greenland Sea and flowing south into the North Atlantic, are well known<sup>12</sup>, but the timing of initiation of deep-water connections between the two ocean basins have been the subject of wide debate in the past two decades<sup>4–9,12</sup>. Settling this debate is essential to refine circulation models that link upwelling of NADW to the onset of Antarctic ice sheet development<sup>2,8,11</sup>. We provide new evidence from seismic-stratigraphic analysis of the Faeroe–Shetland Basin (FSB) that indicates the existence from the early Oligocene of southerly flowing deep-water bottom currents in the axis of this important oceanic gateway.

Previous attempts to constrain the onset of deep-water flow from the Norwegian Sea into the North Atlantic have relied on data from several large, elongated contourite drifts targeted by Deep Sea Drilling Project (DSDP) boreholes southwest of Iceland<sup>12–14</sup> (Fig. 1). Growth of these drifts was most active during the middle Miocene to Pliocene<sup>15</sup>, and this has been used to argue that significant deep-water exchange was established at approximately this time<sup>2,3,8,12,15,16</sup>. A much earlier onset of deep-water connection has been suggested on the basis of a late Eocene date for the basal erosion surface beneath the Feni drift in the Rockall trough and for similar erosion surfaces elsewhere in the North Atlantic<sup>5,10,14,17,18</sup>.

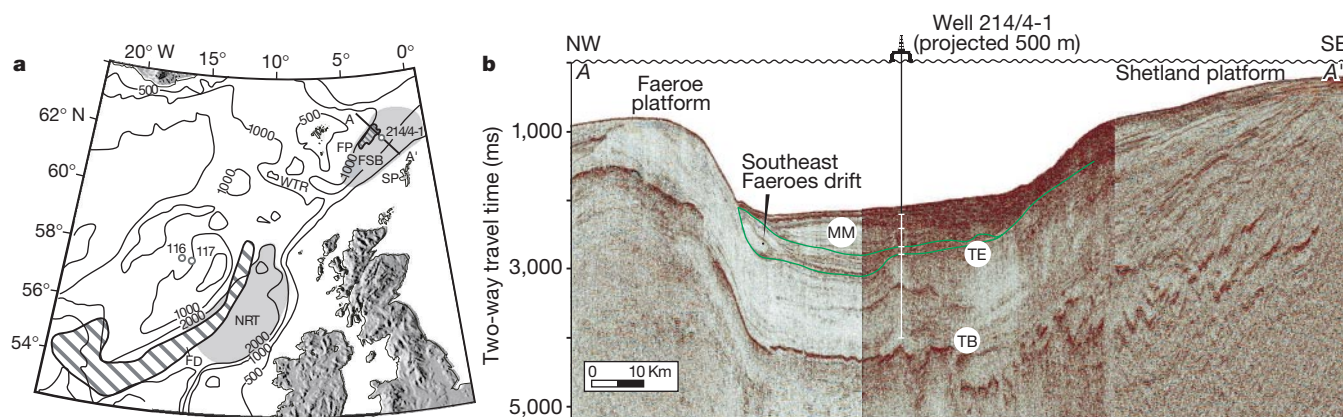
We have identified a previously unrecognized sediment drift immediately northeast of the Greenland–Scotland Ridge (GSR), within the axis of the FSB using a regional grid of more than 15,000 km of multichannel reflection seismic data. The basin fill consists of more than 2 km of Eocene to Recent deep-water sediments<sup>19,20</sup> that was drilled for the first time in mid-1999 (Fig. 1). Preservation of the post-Eocene succession is highly variable within the basin, and is highly attenuated on the south-eastern margin. On the opposing flank (northwest) of the basin, however, there is more complete stratigraphic preservation of this interval<sup>21</sup>, and we have been able to define a large and geometrically distinctive depositional body located at the foot of the southeast Faeroe slope (Fig. 2).

From its external form, internal stratal reflection geometry, and its disposition and stratigraphic context, we interpret this sedimentary feature to be a contourite drift, which we name here the

Southeast Faeroes drift. This interpretation is based on the elongate planform geometry parallel to the basin axis, the common occurrence of moats, and the recognition of an upslope migrating downlap configuration within the body<sup>22,23</sup>. The upslope-facing downlap relationship is a diagnostic geometry of two sub-classes of contourite systems, plastered and separated drifts<sup>23</sup>. In addition to downlap terminations climbing upslope towards the Faeroe margin, we have also identified downlap terminations at the extreme southwestern limit of the drift body within a structurally confined depression where the bottom currents would have been focused (Fig. 2d). These downlapping reflections have a maximum dip of 1–4 degrees in a southwesterly direction, whereas underlying and overlying reflections reveal no sign of post-depositional tilting of that magnitude. This downlap geometry is clearly indicative of the prevailing flow regime<sup>22,23</sup>, and significantly, we conclude that the bottom currents active during deposition of this drift body were flowing in a southwesterly direction.

To date the onset of drift deposition in the Southeast Faeroes drift, we correlated seismic markers a distance of 80 km to a commercial borehole, 214/4-1, for which cuttings were retrieved from the sea bed using a remotely operated vehicle. This was the first time that this sampling technique had been attempted and the well represents a unique data point that allows us to calibrate the stratigraphy of the basin axis, which had thus far proved impossible using industry boreholes located on the Shetland margin. Biostratigraphic, petrographic (scanning electron microscope, SEM) and geochemical analyses of the samples preceding and spanning the interval of drift deposition are summarized in Fig. 3. The seismic horizon defining the base of the drift (marker TE of Fig. 2c) occurs at 2,970 ms (two-way travel time, TWTT) at the borehole location, and is dated with reference to standard biostratigraphic schemes as within the early Oligocene<sup>24,25</sup>. The major erosional surface defining an upper limit for the drift (marker MM of Fig. 2c) is dated as early Pliocene<sup>24,25</sup>.

The seismically defined onset of drift deposition coincides with significant changes in litho- and biostratigraphy that are commensurate with a major change in the circulatory regime in the basin. Sedimentation changed from a mud-dominated clastic system to deposition of biosiliceous ooze (Fig. 3). A prominent high-amplitude event seen at 2,900 ms (TWTT) is calibrated as a porcellanite consisting of more than 70% opal C/T, and represents a diagenetic reflection<sup>26</sup> corresponding to the opal A to C/T boundary. Large changes in the microfauna and flora are also observed across the seismic marker for the ‘base of drift’. Immediately beneath this

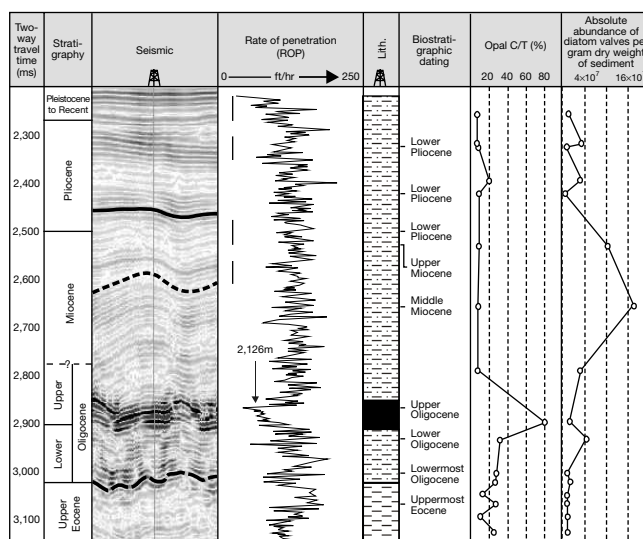


**Figure 1** Location of the study area. **a**, Map showing location of the Faeroe–Shetland basin (FSB) and the Southeast Faeroes drift (elliptical body with diagonal hatching) in relation to the northern Rockall trough (NRT), the Wyville–Thomson ridge (WTR), the Feni drift (FD), the DSDP sites 116, 117, the Shetland platform (SP) and Faeroe platform (FP).

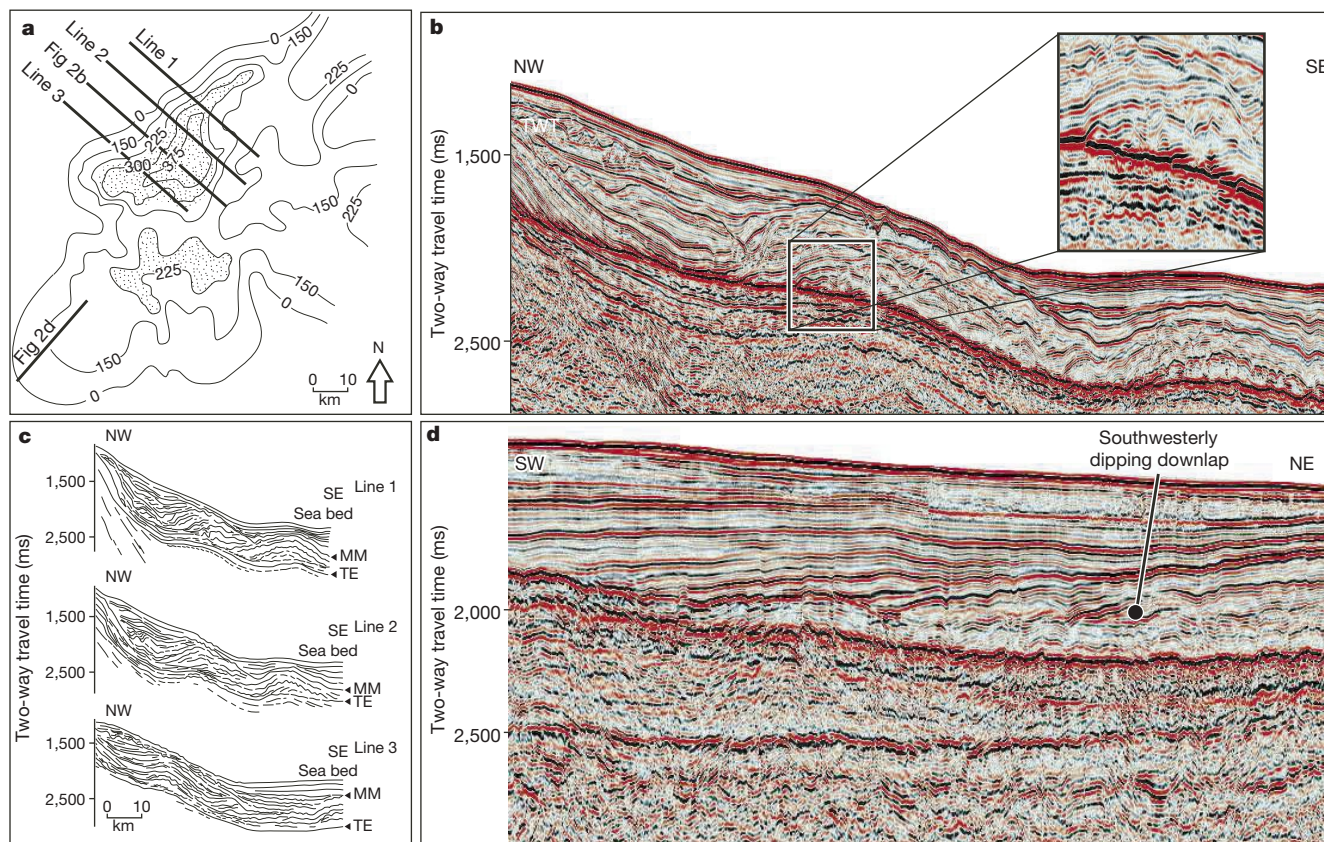
Bathymetric contours are in metres. **b**, Composite multichannel seismic profile across the FSB, showing the location of the 214/4-1 well and the main stratal geometry of the basin fill. Seismic markers include Top Balder (TB), the Top Eocene (TE) and the Middle Miocene (MM).

marker the assemblage consists of more than 90% agglutinated foraminifera, interpreted as due to a poorly oxygenated water column, whereas just above the marker there is an abrupt upward increase in the biogenic silica fraction (diatoms), and a substantial increase in the carbonate fraction expressed as a large influx of calcareous benthic foraminifera, interpreted as due to increasing bottom-water oxygenation. The diatom assemblages contain between 50–70% of near-shore benthic genera such as *Sceptroneis* Ehrenberg and *Paralia* Heiberg; this is an unprecedented observation for deep-sea, basinal sediments, and indicates considerable reworking. This evidence for reworking also supports the inference of a fundamental change in the basin hydrodynamics during the early Oligocene.

The lithological and biostratigraphic changes recognised in 214/4-1 thus independently corroborate the seismic interpretation of the initiation of contourite drift deposition and date this palaeoceanographic event as within the early Oligocene. From the location of the drift and particularly the evidence from the southwesterly dipping downlap geometry, we argue that the change in basin hydrodynamics must have involved the initiation of a deep-water current system flowing to the southwest, along the basin axis in the deepest-water part of the basin, where from clinoform relief we estimate palaeo-water depth to have been over 500 m. Since the onset of drift deposition in the northern Rockall trough (Feni drifts) is dated as late Eocene to early Oligocene<sup>5,13</sup>, and given the similarity between the initial growth of the Feni and the Southeast Faeroes drifts, we conclude that they developed from the same bottom current system.



**Figure 3** Summary of results from the sampled interval of the petroleum industry borehole 214/4-1 and the corresponding well-to-seismic tie (well located on Fig. 1). Vertical scale is two-way time in milliseconds, approximately equivalent to metres. Opal C/T is opal in the form of Cristobalite and Trydymite. Dates used in the text are taken from this well tie, using a consistent geologic timescale<sup>29</sup>.



**Figure 2** Map and seismic geometry of the southeast Faeroe drift. **a**, Time-thickness map of the TE to MM interval over part of the study area (located on Fig. 1), showing northeast–southwest elongated depositional features, parallel to the strike of the depositional slope. Contours indicate two-way travel time in milliseconds. **b**, Seismic profile across the drift body showing the pronounced upslope migrating downlap system, and numerous internal erosion surfaces with moat-like cross-sectional geometry with depths of incision

commonly over 50 m (see enlarged panel). **c**, Line drawings of seismic profiles across the drift exhibiting its diagnostic geometry. The drift base is defined by seismic downlap relationships, stacking in an upslope direction. **d**, Seismic profile along long axis of drift body, showing the critical evidence for a southwesterly flow regime during deposition in the form of southwesterly dipping downlap.



It has been argued that the Feni drift was deposited in a southerly derived counter-clockwise gyre in the Rockall trough, but this was based on the lack of evidence (before the drilling of well 214/4-1) for pre-Neogene (23 Myr ago) drift deposition north of the Wyville–Thomson ridge, and the assumption that the GSR was too shallow to allow deep-water exchange<sup>12</sup>. We suggest that the contemporaneous onset of drift deposition in both basins and the southwesterly flow direction obtained from the Southeast Faeroes drift implies that deep-water exchange must have started within the early Oligocene. The source of the deep water must have been northeast of the Southeast Faeroes drift, and a potential site could have been in a downwelling system in the Norwegian Sea<sup>5</sup>. This locus is consistent with the prevailing cooling in mid–high latitudes in the early Oligocene<sup>9</sup>, and as this northerly deep-water source would have been rich in biosiliceous components and initially poor in calcareous components<sup>9</sup> this would also explain the changes in biota observed in 214/4-1.

We therefore propose that the Southeast Faeroes drift is the earliest expression of NADW flow in the Faeroe–Shetland basin, during the early Oligocene. Its pathway into the open North Atlantic was directed southwest along the basin axis, crossing the Wyville–Thomson ridge into the northern Rockall trough. This is much earlier than the mid–late Miocene timing of deep-water marine connection that is used in current models of North Atlantic palaeocirculation<sup>2,7,11,12,27</sup>. The early onset of NADW formation could have contributed to an interhemispheric feedback mechanism similar to that suggested for the Quaternary<sup>28</sup> that would have helped stabilize Southern-Hemisphere cold climates at the end of the initial Antarctic glacial expansion in the earliest part of the Oligocene. □

Received 15 January; accepted 13 February 2001.

1. Broecker, W. S. The great ocean conveyor. *Oceanography* **4**, 79–89 (1991).
2. Schnitker, D. North Atlantic oceanography as possible cause of Antarctic glaciation and eutrophication. *Nature* **284**, 615–616 (1980).
3. Vogt, P. R. The Faeroe–Iceland–Greenland aseismic ridge and the western boundary undercurrent. *Nature* **239**, 79–81 (1972).
4. Eldholm, O. Paleogene North Atlantic magmatic-tectonic events: environmental implications. *Mem. Soc. Geol. Ital.* **44**, 13–28 (1990).
5. Miller, K. G. & Tucholke, B. E. in *Structure and Development of the Greenland-Scotland Ridge* (eds Bott, M. H. P., Saxov, S., Talwani, M. & Thiede, J.) 549–589 (Plenum, New York, 1983).
6. Thiede, J. & Eldholm, O. in *Structure and Development of the Greenland-Scotland Ridge* (eds Bott, M. H. P., Saxov, S., Talwani, M. & Thiede, J.) 445–455 (Plenum, New York, 1983).
7. Wold, C. N. Cenozoic sediment accumulation on drifts in the northern North Atlantic. *Paleoceanography* **9**, 917–941 (1994).
8. Wright, J. D. & Miller, K. G. Control of North Atlantic Deep Water circulation by the Greenland–Scotland Ridge. *Paleoceanography* **11**, 157–170 (1996).
9. Berggren, W. A. & Schnitker, D. in *Structure and Development of the Greenland-Scotland Ridge* (eds Bott, M. H. P., Saxov, S., Talwani, M. & Thiede, J.) 495–545 (Plenum, New York, 1983).
10. Berggren, W. A. & Hollister, C. D. in *Studies in Paleooceanography* (ed. Hay, W. W.) 126–186 (Society for Economic, Paleontology and Mineralogy, Tulsa, Oklahoma, 1974).
11. Schnitker, D. Global Paleooceanography and its deep water linkage to the Antarctic glaciation. *Earth Sci. Rev.* **16**, 1–20 (1980).
12. Stoker, M. S. in *Geological Processes on Continental Margins* (eds Stoker, M. S., Evans, D. & Cramp, A.) 229–254 (Geological Society, London, 1998).
13. Stow, D. A. V. & Holbrook, J. A. Hatton Drift contourites, northeast Atlantic. *Init. Rep. DSDP* **81**, 695–699 (1984).
14. Kidd, R. B. & Hill, P. R. Sedimentation on Feni and Gardar sediment drifts. *Init. Rep. ODP* **94**, 1217–1244 (1987).
15. Wold, C. N. in *The Tectonics, Sedimentation and Palaeoceanography of the North Atlantic Region* (eds Scrutton, R. A., Stoker, M. S., Shimmield, G. B. & Tudhope, A. W.) 271–302 (Geological Society, London, 1995).
16. Shor, A. N. & Poore, R. Z. Bottom currents and ice rafting in the North Atlantic. *Init. Rep. DSDP* **49**, 859–872 (1979).
17. Nilsen, T. H. in *Structure and Development of the Greenland-Scotland Ridge* (eds Bott, M. H. P., Saxov, S., Talwani, M. & Thiede, J.) 457–477 (Plenum, New York, 1983).
18. Heinrich, R., Wolf, T. C., Bohrmann, G. & Thiede, J. Cenozoic paleoclimatic and paleoceanographic changes in the northern hemisphere revealed by variability of coarse fraction composition in sediments from the Voring Plateau. *Proc. ODP Sci. Res.* **104**, 75–188 (1989).
19. Mitchell, S. M. *et al.* in *Petroleum Geology of Northwest Europe: Proc. 4th Conf.* (ed. Parker, J. R.) 1011–1023 (Geological Society, London, 1993).
20. Davies, R., Cartwright, J. & Rana, J. Giant hummocks in deep-water marine sediments: Evidence for differential compaction and large-scale density inversion during early burial. *Geology* **27**, 907–910 (1999).
21. Boldreel, L. O. & Andersen, M. S. in *The Tectonics, Sedimentation and Palaeoceanography of the North Atlantic Region* (eds Scrutton, R. A., Stoker, M. S., Shimmield, G. B. & Tudhope, A. W.) 145–158 (Geological Society, London, 1995).

22. Hollister, C. D. & Heezen, B. C. in *Studies in Physical Oceanography* (ed. Gordon, A. L.) 37–66 (Gordon and Breach, New York, 1972).
23. Faugères, J.-C., Stow, D. A. V., Imbert, P. & Viana, A. Seismic features diagnostic of contourite drifts. *Mar. Geol.* **162**, 1–38 (1999).
24. Martini, E. Standard Tertiary and Quaternary calcareous nannoplankton zonation. *Proc. 2nd Plankton. Conf. Roma* **2**, 739–785 (1971).
25. Blow, W. H. *The Cainozoic Globigerinida* 1–1413 (Brill, Leiden, 1979).
26. Hein, J. R., Scholl, D. W., Barron, J. A., Jones, M. G. & Miller, J. Diagenesis of late Cenozoic diatomaceous deposits and formation of the bottom-simulating reflector in the southern Bering Sea. *Sedimentology* **25**, 144–181 (1978).
27. Dean, K., McLachlan, K. & Chambers, A. in *Petroleum Geology of Northwest Europe: Proc. 5th Conf.* (eds Fleet, A. J. & Boldy, S. A. R.) 533–544 (Geological Society, London, 1999).
28. Crowley, T. J. North Atlantic Deep Water cools the southern hemisphere. *Paleoceanography* **7**, 489–497 (1992).
29. Harland, W. B. *et al.* *A Geologic Time Scale 1989*. (Cambridge Univ. Press, Cambridge, 1990).

## Acknowledgements

We thank ExxonMobil International Ltd, Phillips Petroleum Company UK Limited, Marathon Oil UK Limited, Statoil (UK) Limited, Enterprise Oil plc, Anadarko North Sea Company, PanCanadian North Sea Limited, BPAmoco, Fugro-Geoteam, PGS Exploration (UK) Ltd and Westerngeco for permission to publish unreleased data. We thank T. Ramsay and C. Harris for discussion of the nannoplanktonic and microfossil record from well 214/4-1.

Correspondence and requests for materials should be addressed to R.D. (e-mail: richard\_davies@email.mobil.com) or J.C. (e-mail: cartwrightja@cardiff.ac.uk).

## Melt retention and segregation beneath mid-ocean ridges

Ulrich H. Faul

Research School of Earth Sciences, The Australian National University, Canberra, ACT 0200, Australia

Geochemical models of melting at mid-ocean ridges—particularly those based on trace elements and uranium-decay-series isotopes—predict that melt segregates from the matrix at very low porosities<sup>1–8</sup>, of order 0.1%. Some of these models also require that the melt ascends rapidly<sup>3,5</sup>. But these predictions appear to conflict with seismic data obtained by the mantle electromagnetic and tomography (MELT) experiment<sup>9</sup>. These data reveal, beneath the East Pacific Rise (at 17°S), a region of low velocities several hundred kilometres wide, which is best explained by the presence of 1–2% melt, distributed on a grain scale in disk-shaped geometries<sup>10</sup>. Here I show that these apparently contradictory constraints can be reconciled by taking into account the geometry and resulting permeability of the intergranular network of melt, together with the changing character of the melt as it ascends. A deep, volatile-rich melt with low viscosity and density is mobile at 0.1% porosity, but basaltic melt only becomes mobile at a porosity above 1%. While the volumetric contribution of the volatile-rich melt to the erupted basalts is small, the isotopic disequilibria (except for radium) generated by porous flow of this melt are preserved if melt transport is rapid at the onset of high-productivity melting. Also, because of incomplete extraction, some melt is retained in a broad zone, consistent with the MELT observations.

Because melt is generated on a grain scale, segregation of melt or fluid from the matrix begins with grain-scale porous flow. The permeability ( $k$ ) of porous rocks is generally calculated with

$$k = \frac{d^2 \phi^n}{C} \quad (1)$$

where  $d$  is the channel spacing (grain size),  $\phi$  is the melt/fluid

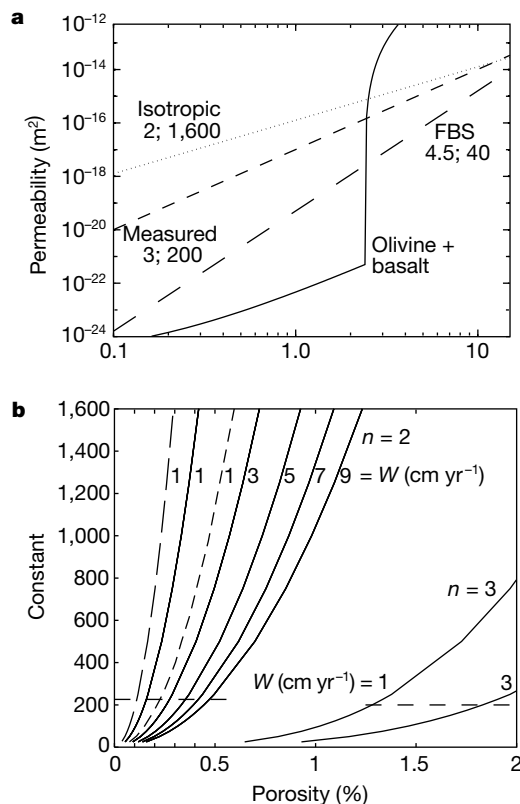


fraction with exponent  $n$  and  $C$  a constant. For porosities below 1%, the permeability is very sensitive to  $n$  (Fig. 1a).  $n = 2$  is calculated for a uniform network of cylinders<sup>11</sup> and an idealized isotropic model of the melt distribution in partially molten rocks<sup>12</sup>. However, for  $n = 2$ , the permeable network has to be self-similar at all porosities, whereas  $n = 3$  is obtained even for a network of cylinders with randomly varying diameters<sup>13</sup>. The relationship between permeability and porosity therefore depends on the behaviour of the network as a whole over a range of porosities; for the self-similar models to be applicable, the pore geometry needs to be close to ideal. Permeability measurements on hot-pressed aggregates of calcite and quartz with water as pore fluid<sup>14</sup>, yielding an exponent

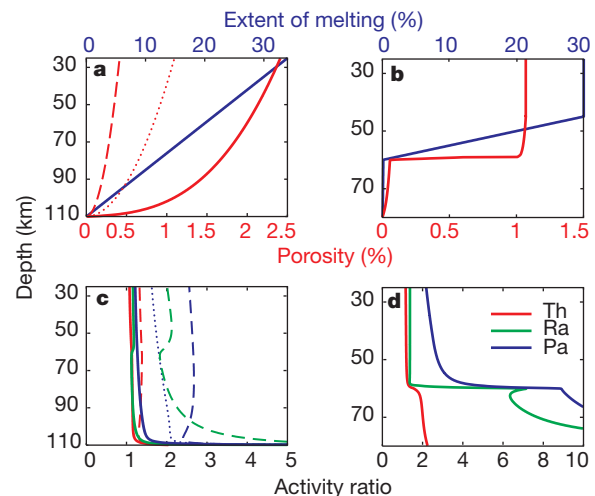
of 3, show that even for texturally equilibrated rocks the pore geometry deviates significantly from the ideal case.

Comparison of the pore geometry in texturally equilibrated calcite with the melt geometry in dunite shows that the pore geometry in calcite is more regular. In calcite, triple-junction tubules account for most of the porosity, but triple-junction tubules contain only about 10% of the total melt content in partially molten dunite<sup>15</sup>. Network modelling<sup>13</sup> emphasizes that the size of the interconnecting throats (the narrowest passages in the pore space) controls the transport properties, even though the contribution of the throats to the overall porosity is less than that of nodal pores. In dunite, most melt resides in geometries best approximated by disks with an axial aspect ratio of 0.05 (ref. 10). The disks are connected by the network of triple-junction tubules, which control the permeability, until the disks become directly interconnected with increasing melt fraction, resulting in a sudden and substantial increase in permeability to values well above the tubule models.

Figure 1a shows a comparison of permeabilities calculated for a self-similar (idealized isotropic) channel network<sup>12</sup>, measured permeabilities<sup>14,16</sup> and experimentally determined melt inclusion shapes<sup>15</sup>. The permeabilities converge at higher porosities, but range over more than six orders of magnitude at 0.1% porosity. These permeabilities can be used to calculate melt-segregation velocities from equation (4) in ref. 17. Assuming the most favourable values for rapid segregation of basaltic melt (viscosity:  $\eta = 1$  Pa s; matrix density – melt density:  $\Delta\rho = 500$  kg m<sup>-3</sup>, values used throughout), segregation velocities for measured permeabilities<sup>14</sup> reach 10 cm yr<sup>-1</sup> at porosities between 1% and 1.5% and 1 m yr<sup>-1</sup> at 4%. Ascent velocities required by dynamic melting models<sup>3,5</sup> are above 1 m yr<sup>-1</sup> at porosities less than 0.5%; this can be achieved either with a melt



**Figure 1** Permeability and resulting maximum porosity in a one-dimensional melting column. **a**, Permeabilities at a grain size of 45  $\mu$ m calculated from a self-similar model (isotropic<sup>12</sup>), measured on texturally equilibrated rocks (measured<sup>14</sup>) and Fontainebleau sandstone (FBS<sup>16</sup>) and predicted for partially molten dunite (olivine+basalt<sup>15</sup>). The exponent and the constant for equation (1) are indicated; the larger the exponent the steeper the slope. For the system olivine plus melt that is below the interconnection threshold for disks, the permeability of the triple-junction tubule network is calculated with  $n = 3$  and  $C = 200$  up to a total melt fraction of 0.024 (melt fraction in triple-junction tubules is about 0.002; see Table 3 in ref. 15). Just below  $\phi = 0.025$  the disks interconnect and the permeability for  $\phi = 0.025$  and 0.033 is calculated from the data and equations given in ref. 15 for disk-shaped inclusions up to an aspect ratio of 0.15. At  $\phi = 0.033$ , nearly all melt inclusions belong to the percolation backbone; the permeability for the disks therefore levels off. The transition from very low permeability to steeply increasing permeability depends on the frequency of wetted two-grain boundaries. For a given porosity this frequency may be grain-size sensitive; it is possible that at mantle grain sizes the transition occurs at somewhat lower porosity than in the experiments. **b**, Maximum porosity at the top of a melting column versus the constant in equation (1) for  $n = 2$  and 3 and a range of solid upwelling velocities  $W$ . Extents of melting are 15% (solid lines) 7.5% (long dashes) and 30% (short dashes). The constants calculated for cylinders on a cubic grid<sup>11</sup> ( $C = 226$ ) and obtained from permeability measurements<sup>14</sup> ( $C = 200$ ) are indicated by horizontal dashed lines. The isotropic model<sup>12</sup> ( $C = 1,600$ , upper limit of **a**) yields porosities  $\geq 0.3\%$  even for slow upwelling and relatively small extents of melting; with measured permeabilities<sup>14</sup> the porosity always exceeds 1%.



**Figure 2** Comparison of isotopic disequilibria produced by different permeabilities at a solid upwelling velocity of 3 cm yr<sup>-1</sup> and with distribution coefficients for an enriched source (Table 1). **a, c**, Standard permeability models have the same melting column as in ref. 4. Only if melt were distributed in cylinders on a cubic grid<sup>11</sup> (dashed lines) could significant excesses be produced by equilibrium porous flow models. With permeabilities measured on texturally equilibrated rocks<sup>14</sup> (solid lines) the porosity very quickly exceeds 1% (constraining the residual porosity in fractional melting models to  $>1\%$ ) and excesses decay within the first kilometre of flow. Th and Ra excesses calculated with the isotropic model (dotted line) are omitted for clarity; they lie below Pa. **b, d**, In the threshold model a small amount of volatile-rich melt is produced 20 km below the garnet–spinel transition, where high-productivity melting begins. Th and Pa excesses are preserved if the melt velocity increases sufficiently rapidly to metres per year. Ra excesses decay within 1 km after the onset of high-productivity melting, where the porosity increases from below 0.1% to above 1%. To preserve significant Ra excesses ( $^{226}\text{Ra}/^{230}\text{Th} \approx 2$ ) the melt needs to enter a fast transport regime ( $\sim 10$  m yr<sup>-1</sup>, requiring 100% melting at  $W = 8$  cm yr<sup>-1</sup> in the one-dimensional model) within 1–2 km of leaving the low porosity regime.

**Table 1 Isotopic excesses calculated with a range of permeability models**

| Permeability                 | MC (km) | F (%) | W (cm yr <sup>-1</sup> ) | $\phi_{\text{max}}$ (%) | <sup>230</sup> Th/ <sup>238</sup> U | <sup>256</sup> Ra/ <sup>230</sup> Th | <sup>231</sup> Pa/ <sup>235</sup> U |
|------------------------------|---------|-------|--------------------------|-------------------------|-------------------------------------|--------------------------------------|-------------------------------------|
| Measured                     | 85, E   | 34    | 3                        | 2.45                    | 1.07                                | 1.16                                 | 1.22                                |
| Idealized isotropic          | 85, E   | 34    | 1/3/5                    | 0.63/1.09/1.41          | 1.23/1.18/1.14                      | 1.56/1.33/1.25                       | 1.91/1.59/1.47                      |
| Cylinders                    | 85, E   | 34    | 3                        | 0.4                     | 1.32                                | 2.01                                 | 2.56                                |
| Cylinders                    | 45, D   | 18    | 3                        | 0.3                     | 1.14                                | 1.71                                 | 1.85                                |
| Threshold                    | 15, E   | 30    | 1/3/5                    | 1.1/1.1/1.1             | 1.10/1.16/1.18                      | 1.37/1.37/1.37                       | 1.54/2.18/2.50                      |
| Threshold                    | 15, D   | 30    | 5                        | 1.1                     | 1.07                                | 1.23                                 | 2.03                                |
| Observed range (mid/maximum) |         |       |                          |                         | 1.1/1.4                             | 2/4                                  | 2/3.6                               |

MC, length of the high-productivity melting column; F, extent of melting; E, 'enriched' source. Bulk distribution coefficients ( $\times 10^{-4}$ ) in the garnet field: U, 45; Th, 20; Pa, 1; in the spinel field: U, 43; Th, 34; Pa, 0.5; Ra, 0.01 throughout. Mineral modes and garnet coefficients are from ref. 5; clinopyroxene coefficients are from ref. 33). D, 'depleted' source (bulk distribution coefficients as in ref. 4). Standard permeability models have the same melting column set-up as ref. 4 (garnet–spinel transition at 60 km depth, cessation of melting at 25 km depth). Threshold models also have the garnet–spinel transition at 60 km depth; volatile-rich melting begins 20 km below, high-productivity melting at the garnet–spinel transition. The steep increase in  $k$  for the threshold models caps the maximum porosity near the threshold porosity.

with the properties of a carbonate melt or with the melt geometry of partially molten lherzolite identical to that of cylinders on a cubic grid.

Uranium-decay-series (<sup>238</sup>U–<sup>230</sup>Th–<sup>226</sup>Ra and <sup>235</sup>U–<sup>231</sup>Pa) disequilibria (excesses of daughter or parent in the erupted melting product resulting from fractionation of parent from daughter due to different melt–solid partition coefficients) measured in MORBs are often used to model mantle melting processes. The model developed by refs 2 and 18 includes the fluid dynamics of melt segregation, but instead of calculating the porosity in the melting column from the relevant material properties ( $k$ ,  $\eta$  and  $\Delta\rho$ ), these properties are implicitly determined by a maximum porosity, which can be chosen to satisfy the U-series data. Here the porosity throughout the melting column is calculated by directly solving equation (16) in ref. 18, with  $k$ ,  $\eta$  and  $\Delta\rho$  as independent input parameters. These calculations show that porosity is not a free parameter in melting models but has well defined physical limits, which together with the mass balance requirement constrain the amount of melt present in the mantle, as well as its ascent velocity.

The maximum porosity at the top of a one-dimensional melting column increases with increasing upwelling velocity and extent of melting (Fig. 1b): calculated isotopic disequilibria will therefore decrease (Table 1). Keeping the maximum porosity constant while varying the matrix upwelling velocity  $W$  (Table 1 in ref. 6) implies that  $k$ ,  $\eta$  and  $\Delta\rho$  vary with matrix upwelling velocity. In order to achieve a maximum porosity near 0.1% (to model Ra excesses<sup>7</sup>), the permeability of partially molten rocks needs to exceed that of cylinders on a cubic grid<sup>11</sup>; this is the most permeable model of a porous rock which can be physically realized.

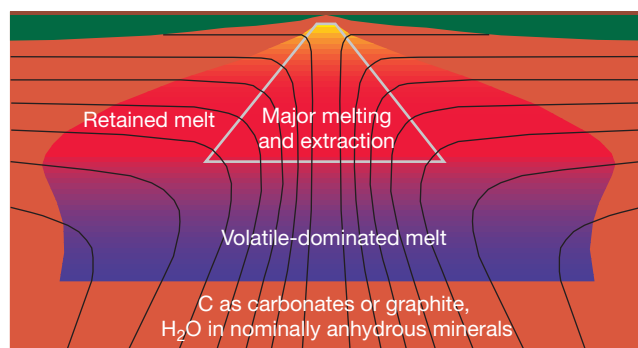
The calculations above show that melt can only segregate at low porosities if its viscosity is substantially lower than that of basaltic

melt. Mid-ocean ridge basalt (MORB) glasses contain between 0.2 and 0.6 wt.% water<sup>19</sup> (regionally averaged) and between 0.2 wt.% (ref. 20) up to 0.7 wt.% CO<sub>2</sub>, equivalent to about 100–400 p.p.m. H<sub>2</sub>O and a similar amount of CO<sub>2</sub> in the MORB source. This suggests that a hydrous fluid with dissolved silica and carbonate components is present at depth below the anhydrous peridotite solidus<sup>21,22</sup>. These C–O–H–silicate fluids or volatile-rich melts become interconnected at all porosities at pressures greater than about 2 GPa (ref. 23). Volatile-dominated melt will have low viscosity and low density, similar to carbonate melt<sup>24</sup>, which is also well interconnected in an olivine matrix<sup>25</sup>. Owing to the incompatibility of trace elements and U-series isotopes in the crystalline matrix, they will preferentially partition into any fluid or melt phase. The potential importance of a mobile, volatile-rich melt has been discussed before<sup>17,26–28</sup>; however, basaltic melt was also predicted to segregate rapidly at porosities below 1% (refs 17, 28).

The strong dependence of the calculated isotopic disequilibria on melt and matrix properties is illustrated in Fig. 2. In model calculations essentially identical to those of refs 4, 6 and 7 (Fig. 2a and c) significant excesses can only be produced if all melt is contained in uniformly varying cylinders on a cubic grid. With measured permeabilities the excesses decay soon after the first melt is formed, as the porosity rapidly exceeds 1%. With a volatile-rich melt at very low porosities and a rapid increase in permeability above a threshold, Th and Pa excesses can be preserved, but require rapid melt production at the onset of high-productivity melting (Fig. 2b and d). Fast melt production could be due to melting of inhomogeneities (for example, pyroxenites<sup>29</sup>) or melt focusing from two or three dimensions<sup>30</sup>. Figure 2d also shows that it is not possible to generate significant Ra excesses by dynamic melting<sup>3</sup>. Ra excesses are only produced at very low porosities, and although in this model the melt velocity increases rapidly to several metres per year as porosity increases from ~0.1% to ~1%, Ra excesses decay already at the onset of this transition regime (see also ref. 7). It is therefore not possible to generate substantial Ra excesses by conventional mantle melting models, using physically reasonable parameters.

Melt also moves by porous flow in dunite channels, so the same constraints on permeability apply. Owing to the steep increase of the permeability in the threshold model, which caps the porosity near the threshold, the porosity in channels will be only slightly higher than in the adjacent areas.

The above discussion suggests that Ra excesses are produced in the shallow mantle. One possibility is the involvement of mantle exposed at the ocean floor (abyssal peridotite). The high Cl contents observed in ref. 31 in some melt inclusions in plagioclase in MORB may be the result of re-infiltration of melt into relatively cool mantle exposed to hydrothermal circulation. As only very few samples with measured high Ra excesses also have reported Cl analyses, this origin of Ra excesses is a hypothesis at present. However, the two samples with high Ra excess in ref. 4 for which Cl analysis are available have high Cl contents; the four samples with some of the largest



**Figure 3** Sketch of the distribution of melt beneath a mid-ocean ridge. The black lines indicate matrix flow lines for a corner flow model. Because melt extraction takes place above 1% porosity, some melt is retained when the matrix flow lines turn horizontal, resulting in the broad low velocity zone imaged by the MELT experiment. The region from which melt is extracted is much smaller (see for example ref. 27).

Ra excesses measured in MORB to date<sup>8</sup> all come from a graben in the Siqueiros Transform, which has undergone relatively recent extension.

The effect of variable viscosity and density on melt segregation beneath mid-ocean ridges is shown schematically in Fig. 3. A small amount of volatile-rich melt is formed in the upwelling part of the mantle by decarbonation reactions or by oxidation if carbon is stored as graphite or diamond. As water is incompatible in nominally anhydrous minerals relative to melt, it will partition into the melt, although the release might be gradual<sup>32</sup>. This volatile-rich melt is able to move at small porosities because of its low viscosity and density. Because the contribution of this melt to the volume of erupted basalts is small, its incompatible element-enriched character is lost, but isotopic disequilibria are not affected by dilution. Owing to incomplete extraction, some melt (more than 1%) remains in a broad zone, consistent with the observations from the MELT experiment. A two-stage melting model could also explain the depleted character of clinopyroxene in abyssal peridotites<sup>1</sup>. Modelling<sup>28</sup> showed that the observed rare earth element patterns require low porosities only at one stage but not throughout melt extraction. □

Received 15 September 2000; accepted 12 February 2001.

- Johnson, K. H., Dick, H. J. B. & Shimizu, N. Melting in the oceanic upper mantle: An ion microprobe study of diopsides in abyssal peridotites. *J. Geophys. Res.* **95**, 2661–2678 (1990).
- Spiegelman, M. & Elliott, T. Consequences of melt transport for uranium series disequilibrium in young lavas. *Earth Planet. Sci. Lett.* **118**, 1–20 (1993).
- Richardson, C. & McKenzie, D. Radioactive disequilibria from 2D models of melt generation by plumes and ridges. *Earth Planet. Sci. Lett.* **128**, 425–437 (1994).
- Lundstrom, C. C., Gill, J., Williams, Q. & Perfit, M. R. Mantle melting and basalt extraction by equilibrium porous flow. *Science* **270**, 1958–1961 (1995).
- Salter, V. J. M. & Longhi, J. Trace element partitioning during the initial stages of melting beneath mid-ocean ridges. *Earth Planet. Sci. Lett.* **166**, 15–30 (1999).
- Lundstrom, C. C., Williams, Q. & Gill, J. B. Investigating solid mantle upwelling rates beneath mid-ocean ridges using U-series disequilibria, 1: a global approach. *Earth Planet. Sci. Lett.* **157**, 151–165 (1998).
- Lundstrom, C. C., Gill, J. & Williams, Q. A geochemically consistent hypothesis for MORB generation. *Chem. Geol.* **162**, 105–126 (2000).
- Lundstrom, C. C., Sampson, D. E., Perfit, M. R., Gill, J. & Williams, Q. Insights into mid-ocean ridge basalt petrogenesis: U-series disequilibria from the Siqueiros Transform, Lamont Seamounts, and East Pacific Rise. *J. Geophys. Res.* **104**, 13035–13048 (1999).
- The MELT Seismic Team. Imaging the deep seismic structure beneath a mid-ocean ridge: the MELT experiment. *Science* **280**, 1215–1218 (1998).
- Faul, U. H., Toomey, D. R. & Waff, H. S. Intergranular basaltic melt is distributed in thin, elongated inclusions. *Geophys. Res. Lett.* **21**, 29–32 (1994).
- Turcotte, D. L. & Schubert, G. *Geodynamics* (John Wiley, New York, 1982).
- von Bargen, N. & Waff, H. S. Permeabilities, interfacial areas and curvatures of partially molten systems: Results of numerical computations of equilibrium microstructures. *J. Geophys. Res.* **91**, 9261–9276 (1986).
- Zhu, W., David, C. & Wong, T.-f. Network modeling of permeability evolution during cementation and hot isostatic pressing. *J. Geophys. Res.* **100**, 15451–15464 (1995).
- Wark, D. A. & Watson, E. B. Grain-scale permeabilities of texturally equilibrated, monomineralic rocks. *Earth Planet. Sci. Lett.* **164**, 591–605 (1998).
- Faul, U. H. Permeability of partially molten upper mantle rocks from experiments and percolation theory. *J. Geophys. Res.* **102**, 10299–10311 (1997).
- Bourbie, T. & Zinsner, B. Hydraulic and acoustic properties as a function of porosity in Fontainebleau sandstone. *J. Geophys. Res.* **90**, 11524–11532 (1985).
- McKenzie, D. Some remarks on the movement of small melt fractions in the mantle. *Earth Planet. Sci. Lett.* **95**, 53–72 (1989).
- Spiegelman, M. UserCalc: A Web-based uranium series calculator for magma migration problems. *Geochim. Geophys. Geosyst.* **1**, 1999 GC 0000 30 (2000).
- Michael, P. Regionally distinctive sources of depleted MORB: Evidence from trace elements and H<sub>2</sub>O. *Earth Planet. Sci. Lett.* **131**, 301–320 (1995).
- Zhang, Y. & Zindler, A. Distribution and evolution of carbon and nitrogen in the Earth. *Earth Planet. Sci. Lett.* **117**, 331–345 (1993).
- Green, D. H. & Liebermann, R. C. Phase equilibria and elastic properties of a pyrolite model for the oceanic upper mantle. *Tectonophysics* **32**, 61–92 (1976).
- Blundy, J. D., Brodholt, J. P. & Wood, B. J. Carbon-fluid equilibria and the oxidation state of the upper mantle. *Nature* **349**, 321–324 (1991).
- Mibe, K., Fujii, T. & Yasuda, A. Connectivity of aqueous fluid in the Earth's upper mantle. *Geophys. Res. Lett.* **25**, 1233–1236 (1998).
- Dobson, D. P. et al. In-situ measurement of viscosity and density of carbonate melts at high pressure. *Earth Planet. Sci. Lett.* **143**, 207–215 (1996).
- Hunter, R. H. & McKenzie, D. The equilibrium geometry of carbonate melts in rocks of mantle composition. *Earth Planet. Sci. Lett.* **92**, 347–356 (1989).
- Galer, S. J. G. & O'Nions, R. K. Magmagenesis and the mapping of chemical and isotopic variations in the mantle. *Chem. Geol.* **56**, 45–61 (1985).
- Plank, T. & Langmuir, C. H. Effects of the melting regime on the composition of the oceanic crust. *J. Geophys. Res.* **97**, 19749–19770 (1992).

- Kelemen, P. B., Hirth, G., Shimizu, N., Spiegelman, M. & Dick, H. J. B. A review of melt migration processes in the adiabatically upwelling mantle beneath oceanic spreading ridges. *Phil. Trans. R. Soc. Lond. A* **355**, 283–318 (1997).
- Yaxley, G. M. Experimental study of the phase and melting relations of homogeneous basalt + peridotite mixtures and implications for the petrogenesis of flood basalts. *Contrib. Mineral. Petrol.* **139**, 326–338 (2000).
- Asimow, P. D. & Stolper, E. M. Steady-state mantle–melt interactions in one dimension: I. Equilibrium transport and melt focusing. *J. Petrol.* **40**, 475–494 (1999).
- Nielsen, R. L., Sours-Page, R. E. & Harpp, K. S. Role of a Cl-bearing flux in the origin of depleted ocean floor magmas. *Geochim. Geophys. Geosyst.* **1**, 1999 GC 0000 17 (2000).
- Hirth, G. & Kohlstedt, D. L. Water in the oceanic upper mantle: implications for rheology, melt extraction and evolution of the lithosphere. *Earth Planet. Sci. Lett.* **144**, 93–108 (1996).
- Wood, B. J., Blundy, J. D. & Robinson, J. A. C. The role of clinopyroxene in generating U-series disequilibrium during mantle melting. *Geochim. Cosmochim. Acta* **63**, 1613–1620 (1999).

## Acknowledgements

Discussions with S. Eggins, D. Green and H. O'Neill are gratefully acknowledged, as well as pre-publication access to a G<sup>3</sup> article by M. Spiegelman. Reviews by T. Elliot and M. Spiegelman helped to shape this manuscript.

Correspondence and requests for materials should be addressed to the author (e-mail: uli.faul@anu.edu.au).

# Towards a general theory of biodiversity

Elizaveta Pachepsky\*†, John W. Crawford†, James L. Bown† & Geoff Squire\*

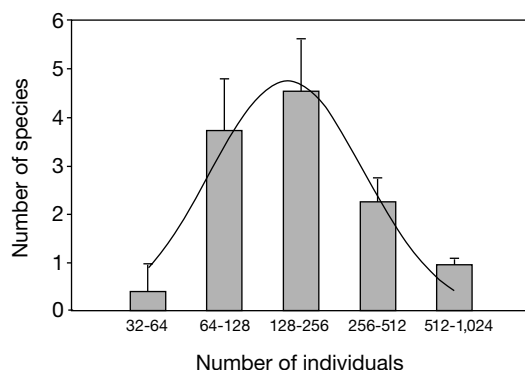
† SIMBIOS Centre, School of Computing, University of Abertay Dundee, Dundee DD1 1HG, UK

\* Vegetation Systems Unit, Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, UK

The study of patterns in living diversity is driven by the desire to find the universal rules that underlie the organization of ecosystems<sup>1,2</sup>. The relative abundance distribution, which characterizes the total number and abundance of species in a community, is arguably the most fundamental measure in ecology. Considerable effort has been expended in striving for a general theory that can explain the form of the distribution<sup>3,4</sup>. Despite this, a mechanistic understanding of the form in terms of physiological and environmental parameters remains elusive<sup>5</sup>. Recently, it has been proposed that space plays a central role in generating the patterns of diversity<sup>6,7</sup>. Here we show that an understanding of the observed form of the relative abundance distribution requires a consideration of how individuals pack in time. We present a framework for studying the dynamics of communities which generalizes the prevailing species-based approach to one based on individuals that are characterized by their physiological traits. The observed form of the abundance distribution and its dependence on richness and disturbance are reproduced, and can be understood in terms of the trade-off between time to reproduction and fecundity.

The relative abundance distribution describes how the individuals in a community are partitioned among rare and common species. A log-normal form of the distribution, implying a community containing many rare species and relatively few common ones, is associated with a community in equilibrium. A power-law (or geometric) form, implying a more equitable share of individuals amongst species groups, is associated with non-equilibrium behaviour resulting from perturbation due to disturbance, pollution or immigration<sup>8</sup>. Two pervasive theories—multiplicative recruitment<sup>3</sup> and sequential niche breakage<sup>4</sup>—attempt to explain the emergent log-normal form of the abundance distribution in general terms, but neither of these provides an explanation of the observed



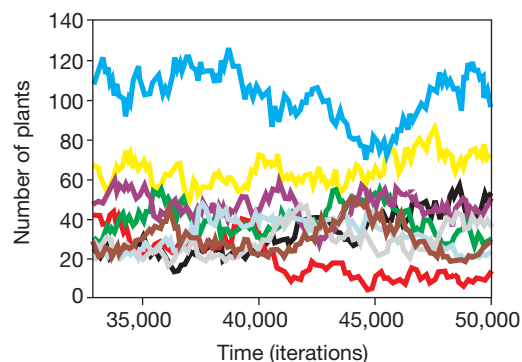


**Figure 1** Sample abundance distribution on a  $30 \times 30$  lattice. The histogram values were averaged over the time of coexistence of 15 plant types, and the error bars (corresponding

to the standard deviation) give an indication of the variation in time within the abundance classes.

patterns<sup>5</sup>. An explanation in terms of immigration, migration and speciation at the metapopulation scale has been developed<sup>9</sup>, but is phenomenological and does not address the origins of relative abundance at the population scale. Fundamentally, few models employ parameters that can be measured directly in individuals, and almost all average out the substantial variation that exists between individuals of the same species. Therefore, the functional properties of individuals and the resulting state of the community cannot be mechanistically linked, and this fuels the topical debate on the relation between biological diversity and community function<sup>10</sup>.

To address these issues, we have developed an individual-based model with three main features. (1) Individual-based—a plant community is modelled by explicitly simulating development of each individual plant in terms of physiological traits. The spread of values of the plants' traits forms a probability distribution of trait values that defines the diversity in a community. This departs from the usual definition of diversity based on species, and leads to a definition in terms of the variables governing the dynamics of the community. (2) Spatially explicit—individuals are located at nodes on a discrete spatial lattice, and interact within a neighbourhood that depends on each plant's development stage. Only one plant can exist at each lattice location. Nutrients are depleted from the resource base in accordance with plants' uptake. The resource base is distributed uniformly in space, because the focus of the work reported here is on the role of individual behaviour in promoting diversity. (The effect of environmental heterogeneity on diversity is well documented<sup>6,11</sup>.) We assume that the lattice represents an isolated community<sup>6</sup>; that is, we exclude the effects of immigration, and seeds that land outside the lattice do not contribute to the dynamics. (3) Process-based—a crucial feature of the approach is that the traits represent physiological processes characterized by experimentally derived parameters. An individual is defined in terms of 12 traits that relate to resource uptake, area over which resource is captured, internal allocation of resources between structure, storage and reproduction, time of reproduction, number of progeny produced, dispersal of progeny, and survival. Traits are primarily functions of a plant's development state. Competition in the model occurs for resource and for space. The resource competition occurs when more than one plant takes up resource from the same location on the lattice, in which case the resource is distributed in proportion to plants' current resource uptake. Competition for space occurs when seeds are distributed on the lattice. A seed germinates only if it falls on an empty lattice location. In the results reported here, dispersal area is constant for all plants, and the seeds are dispersed randomly. An analysis of the effect of varying the dispersal parameters showed that diversity increases as dispersal is progressively reduced below the level of complete mixing, and was consistent with the results of previous studies<sup>7,12–14</sup>. To eliminate

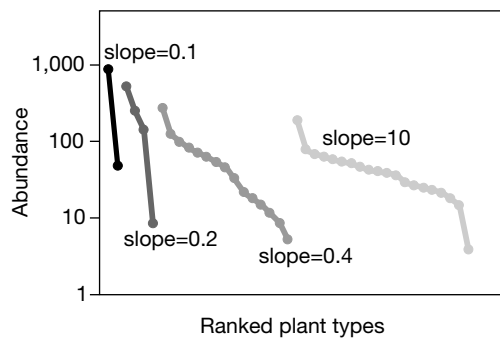


**Figure 2** Evolution of abundance of nine plant types in time on a  $20 \times 20$  lattice.

gene flow as a cause of any observed pattern, individuals reproduce clonally in this study. The full set of parameters and the corresponding probability density functions were derived from experiments conducted on the grassland species *Rumex acetosa* L. as part of a wider study of the dynamics of species-rich grasslands (U. Bausenwein *et al.*, manuscript in preparation).

The model was initialized with 75 plants randomly located on the lattice, and with trait value sets randomly selected from the trait probability density functions. The size of the lattices used range from  $10 \times 10$  to  $50 \times 50$  (equivalent to approximately  $1 \text{ m}^2$  to  $25 \text{ m}^2$  for *R. acetosa*). The model was run for 50,000 iterations, corresponding to approximately 1,250 generations. In the initial stages, the number of types decreased exponentially with time, such that only half the number of types remained after an average of 5,000 iterations. Subsequently, a dynamic equilibrium was achieved with the number of types remaining constant for at least 1,000 generations. The number of coexisting types depended on the area of the lattice. Across the range of scales studied, the dependence on area was of a power-law form with exponent equal to  $0.4 \pm 0.06$ , consistent with the range observed in real ecological communities<sup>15</sup>.

The resulting abundance distribution was of log-normal form for the cases where the system had been in equilibrium for the longest time, and where a sufficiently large number of individuals remained at the end of the exponential phase for meaningful statistics ( $P > 0.2$  for the Kolmogorov–Smirnov test with Lilliefors corrected critical values<sup>16</sup>). An example is shown in Fig. 1. During the exponential phase, the abundance distribution was more similar to the geometric form observed for disturbed or non-equilibrium communities. Not only is the log-normal form of the species relative abundance distribution reproduced, but the correspondence between the standard deviation of the



**Figure 3** Effect of the trade-off between time of reproduction and fecundity on the shape of the abundance curve on a  $50 \times 50$  lattice. Results were obtained from a simplified version of the individual-based model with fecundity = slope  $\times$  (time of reproduction) + 1.5. Individual mortality is random, with a probability of 0.001 per cycle.

distribution and richness (that is, the total number of types) is the same as that observed in communities of less than a few hundred species<sup>4</sup>. The parameters are also stable over time, despite the fact that the ranks of plant types are not constant. In all cases, there was clear stratification in the abundance classes (Fig. 2). In particular, we observed that the most abundant individuals in the community also had the most stable numbers. The ranks within the least abundant classes appeared to be governed by stochastic processes relating to competition for the space vacated through death of a resident. This dynamic behaviour of abundance has been observed in communities of ground beetles<sup>17</sup> and plankton<sup>18</sup>, and illustrates the underlying dynamic form of the equilibrium state. Although the trait values of the modelled population are based on ranges observed in a particular species, *R. acetosa*, the model produces general patterns that are common to various ecological communities.

In order to identify the aspects of individual variation that were required to generate the observed diversity patterns, the effect of the variation of each of the physiological traits was studied. This was done by replacing the probability distribution by the mean value for each trait in turn, and looking at the effect on the diversity in the simulated communities. Of the 12 traits in the model, only variation in those affecting time to reproduction and fecundity was found to influence the abundance of individuals in the community. A plot of time to reproduction against fecundity for the coexisting individuals revealed an approximately linear trade-off. This observation allowed us to reduce the complexity of the model from a form using 12 traits to one using only three: time of reproduction, a linear relation (trade-off) between time of reproduction and fecundity, and a small random death factor (see Fig. 3). In the simplified model, a lattice is populated by individuals that vary in their time to reproduction and produce an amount of seed in accordance with the trade-off, disperse this seed as in the full model, and die at random. The initial conditions, the lattice size and the duration of the simulation were as described above. This model exhibited the same underlying dynamics, abundance distribution and species-area curve to that observed in the full model. Furthermore, simulations showed that the parameters of the predicted abundance distribution were sensitive to the form of the trade-off (Fig. 3). The important environmental factor in the model was weak stochastic disturbance in the form of random death of a small percentage of individuals. Without this, richness was significantly reduced.

The generality of these conclusions can be established using a mathematical model that incorporates the basic processes governing the dynamics of most ecological communities: competition for resource, reproduction and death. The local interactions between individuals are ignored in this model, and individuals are assumed

# Box 1

## Three-dimensional trait-space model

In the mean-field model, we assume that four processes define an individual in terms of its contribution to the dynamics of the community: birth, competition, reproduction and death. Thus an individual is characterized by the time to reproduction,  $1/\lambda$ ; lifetime,  $1/a$ ; the number of progeny produced per unit time,  $b$ ; and the competition that arises through density-dependent interactions between individuals. The model describes the effect of competition on the rate of conversion of non-reproductive juveniles,  $s_{\lambda,a,b}$ , into reproductive adults,  $f_{\lambda,a,b}$ . The diversity of the community is described by the frequency distribution of  $f_{\lambda,a,b}$  and  $s_{\lambda,a,b}$  across the trait values:

$$\frac{\partial f_{\lambda,a,b}(t)}{\partial t} = \frac{\lambda s_{\lambda,a,b}(t)}{1 + K \int_0^{\lambda_c} \kappa(x - \lambda) s_{x,a,b}(t) dx} - a f_{\lambda,a,b}(t)$$

$$\frac{\partial s_{\lambda,a,b}(t)}{\partial t} = b \int_{\Omega_x} \int_{\Omega_y} p(\mathbf{u}|\mathbf{x}, \mathbf{y}) G[f_x, f_y] dx dy - c \lambda s_{\lambda,a,b}(t)$$

The single integral term arises from the assumption that competition between individuals depends on their relative times to reproduction through the form of  $\kappa$ . Both  $K$  and  $c$  are assumed constant, and  $1/\lambda_c$  is the minimum time to reproduction. Gene flow is expressed through the function  $p(\mathbf{u}|\mathbf{x}, \mathbf{y})$  that describes the probability that an individual of type  $\mathbf{u} = \{\lambda, a, b\}$  results from a cross between individuals of types  $\mathbf{x} = \{\lambda_x, a_x, b_x\}$  and  $\mathbf{y} = \{\lambda_y, a_y, b_y\}$ , which are two points in the three-dimensional space of the parameters  $\lambda$ ,  $a$  and  $b$ . The function  $G[f_x, f_y]$  describes the frequency of crossing between individuals of type  $\mathbf{x}$  and  $\mathbf{y}$ , and the integral is over the parameter space,  $\Omega_x$  and  $\Omega_y$ .

The analytical solutions of this model quoted in the text refer to the case where we have assumed for simplicity that  $\kappa$  can be approximated by the Heaviside function,  $u_\lambda(x)$  (that is,  $u_\lambda(x) = 0$  for  $x < \lambda$  and  $u_\lambda(x) = 1$  for  $x \geq \lambda$ ). The special case of clonal reproduction is accommodated by replacing the double integral term by the term  $b f_{\lambda,a,b}(t)$ . Therefore, we consider solutions to the system:

$$\frac{\partial f_{\lambda,a,b}(t)}{\partial t} = \frac{\lambda s_{\lambda,a,b}(t)}{1 + K \int_0^{\lambda_c} s_{x,a,b}(t) dx} - a f_{\lambda,a,b}(t)$$

$$\frac{\partial s_{\lambda,a,b}(t)}{\partial t} = b f_{\lambda,a,b}(t) - c \lambda s_{\lambda,a,b}(t)$$

The steady-state solutions of this system require that  $b/a$  is a function of  $\lambda$ , otherwise the distribution along  $\lambda$  collapses to a point at  $\lambda_c$ . The steady-state solutions are

$$s_{\lambda,a,b}(t) = -\frac{1}{cK} \frac{d}{d\lambda} \left( \frac{b(\lambda)}{a(\lambda)} \right) \quad (1)$$

$$f_{\lambda,a,b}(t) = -\frac{\lambda}{Kb(\lambda)} \frac{d}{d\lambda} \left( \frac{b(\lambda)}{a(\lambda)} \right) \quad (2)$$

with an additional condition that  $b(1)/a(1) = 1$ .

From (1) it follows that  $d[b(\lambda)/a(\lambda)]/d\lambda < 0$ , that is,  $b(\lambda)/a(\lambda)$  must be monotonic decreasing. Moreover, the linear stability analysis shows that for non-zero steady state, it is necessary that  $d^2[b(\lambda)/a(\lambda)]/d\lambda^2 > 0$ , that is, that  $b(\lambda)/a(\lambda)$  is concave up. The non-zero steady state is stable if  $b(\lambda)/a(\lambda) > cK$ , and otherwise the 0-state is stable.

Using heuristic forms for  $\mathbf{p}$  and  $\mathbf{G}$ , it can be shown that the effect of trade-offs on the abundance curve can be modified by gene flow through the form of the dependence of the product of  $\mathbf{p}$  and  $\mathbf{G}$  on the parameters.

to be spatially well mixed. The dynamics of the community can then be described by two ordinary differential equations (Box 1) describing the time evolution of the system in a three-dimensional trait space. The traits that characterize the above processes are the time to reproduction ( $1/\lambda$ ), the characteristic lifespan ( $1/a$ ), and fecundity ( $b$ ). The solution shows that unless the ratio  $b/a$  is a function of  $\lambda$ , the system collapses to a state where only the fastest

reproducing individuals remain. Therefore, an interdependence between time to reproduction and at least one of longevity or fecundity is required to maintain diversity. The form of the trade-off determines the shape of the distribution of abundance of coexisting individuals across the trait space and the stability of that distribution (Box 1).

These results confirm the main features of the individual-based model—that the trade-off between time to reproduction and fecundity sustains the diversity in the community, and governs the form of the resulting abundance distribution. The trade-off is a simple but fundamental example of the link between individual- and community-scale features of the system. Its central role suggests that an explanation of community dynamics must incorporate an account of individual behaviour. Furthermore, by integrating more complex mechanisms encompassing uptake, competition and the effect of the nutritional status of an individual on its fecundity, the trade-off is also more amenable to measurement than the plethora of underlying finer-scale mechanisms. Although the differential equation model reproduced the qualitative results of the individual-based model, the predicted abundance distribution is not log-normal (except for very special choices of the form of the trade-off). An explicit account of space, as in the individual-based model, seems to be required to produce the log-normal form of the abundance distribution at equilibrium. Neither geneflow nor mutation are essential processes in the generation or maintenance of the log-normal form, although they are likely to modify the associated parameters (Box 1).

We have shown that significant diversity in traits can be sustained on small spatial domains, and that this diversity possesses the observed forms of community-scale patterns. This suggests that progress can be made using relatively small systems which are amenable to experimentation, and replacing species with individuals as the fundamental ecological accounting unit. □

Received 12 October 2000; accepted 18 January 2001.

1. Sole, R. V., Manrubia, S. C., Benton, M., Kauffman, S. & Bak, P. Criticality and scaling in evolutionary ecology. *Trends Ecol. Evol.* **14**, 156–160 (1999).
2. West, G. B., Brown, J. H. & Enquist, B. J. A general model for the structure and allometry of plant vascular systems. *Nature* **400**, 664–667 (1999).
3. MacArthur, R. M. On the relative abundance of species. *Am. Nat.* **94**, 25–36 (1960).
4. Sugihara, G. Minimal community structure: an explanation of species abundance patterns. *Am. Nat.* **116**, 770–787 (1980).
5. Ueland, K. I. & Gray, J. S. Lognormal distributions and the concept of community equilibrium. *Oikos* **39**, 171–178 (1982).
6. Tilman, D. & Kareiva, P. *Spatial Ecology* (Princeton Univ. Press, Princeton, 1997).
7. Giplin, M. & Hanski, I. *Metapopulation Dynamics: Empirical and Theoretical Investigations* (Academic, London, 1991).
8. Putman, R. J. *Community Ecology* (Chapman and Hall, London, 1994).
9. Hubbell, S. P. A unified theory of biogeography and relative species abundance and its application to tropical rain forests and coral reefs. *Coral Reefs* **16** (suppl.), S9–S21 (1997).
10. Huston, M. A. *Biological Diversity: The Coexistence of Species on Changing Landscapes* (Cambridge Univ. Press, Cambridge, 1994).
11. Durrett, R. & Levin, S. A. The importance of being discrete (and spatial). *Theor. Popul. Biol.* **46**, 363–394 (1994).
12. Winkler, E., Fischer, M. & Schmid, B. Modelling the competitiveness of clonal plants by complementary analytical and simulation approaches. *Oikos* **85**, 217–233 (1999).
13. Pacala, S. W. Neighbouring models of plant population dynamics. 2. Multi-species models of annuals. *Theor. Popul. Biol.* **29**, 262–292 (1986).
14. Crawley, M. J. & May, R. M. Population dynamics and plant community structure: competition between annuals and perennials. *J. Theor. Biol.* **125**, 475–489 (1987).
15. Rosenzweig, M. L. *Species Diversity in Space and Time* (Cambridge Univ. Press, Cambridge, 1995).
16. Legendre, L. & Legendre, P. *Numerical Ecology* (Elsevier, Amsterdam, 1983).
17. Loreau, M. Species abundance patterns and the structure of ground-beetle communities. *Ann. Zool. Fennici* **28**, 49–56 (1992).
18. Pearson, T. H., Gray, J. S. & Johannessen, P. J. Objective selection of sensitive species indicative of pollution-induced change in benthic communities. 2. Data analyses. *Mar. Ecol. Prog. Ser.* **12**, 237–255 (1983).

## Acknowledgements

We thank U. Bausenwein and P. Millard for supplying the *Rumex acetosa* data used in the model. This work was partly funded by SERAD.

Correspondence and requests for materials should be addressed to J.W.C. (e-mail: j.crawford@tay.ac.uk).

# Sustainability of three apple production systems

John P. Reganold\*, Jerry D. Glover\*, Preston K. Andrews† & Herbert R. Hinman‡

\* Department of Crop and Soil Sciences, Washington State University, Pullman, Washington 99164, USA

† Department of Horticulture and Landscape Architecture, Washington State University, Pullman, Washington 99164, USA

‡ Department of Agricultural Economics, Washington State University, Pullman, Washington 99164, USA

Escalating production costs, heavy reliance on non-renewable resources, reduced biodiversity, water contamination, chemical residues in food, soil degradation and health risks to farm workers handling pesticides all bring into question the sustainability of conventional farming systems<sup>1–4</sup>. It has been claimed<sup>5,6</sup>, however, that organic farming systems are less efficient, pose greater health risks and produce half the yields of conventional farming systems. Nevertheless, organic farming became one of the fastest growing segments of US and European agriculture during the 1990s<sup>7,8</sup>. Integrated farming, using a combination of organic and conventional techniques, has been successfully adopted on a wide scale in Europe<sup>9</sup>. Here we report the sustainability of organic, conventional and integrated apple production systems in Washington State from 1994 to 1999. All three systems gave similar apple yields. The organic and integrated systems had higher soil quality and potentially lower negative environmental impact than the conventional system. When compared with the conventional and integrated systems, the organic system produced sweeter and less tart apples, higher profitability and greater energy efficiency. Our data indicate that the organic system ranked first in environmental and economic sustainability, the integrated system second and the conventional system last.

Organic management practices combine traditional conservation-minded farming methods with modern farming technologies but exclude such conventional inputs as synthetic pesticides and fertilizers, instead putting the emphasis on building up the soil with compost additions and animal and green manures, controlling pests naturally, rotating crops and diversifying crops and livestock<sup>10</sup>. Organic farming systems in the US range from strict closed-cycle systems that go beyond organic certification guidelines by limiting, as much as possible, external inputs to more standard systems that simply follow organic certification guidelines. Integrated farming systems reduce the use of chemicals by integrating organic and conventional production methods.

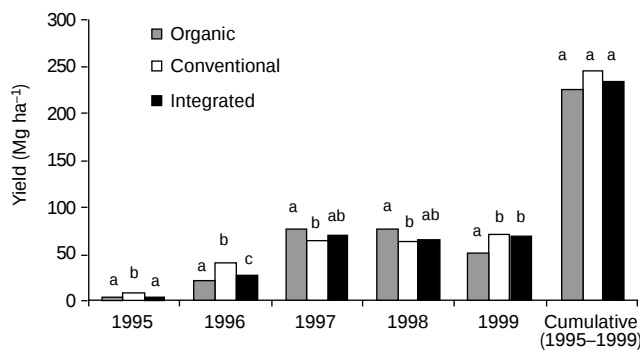
Just because a system is organic or integrated does not ensure its sustainability; nor does sustainability, an inherently complex

**Table 1 Soil quality ratings of three apple production systems**

| Soil quality functions                     | Year | Organic | Conventional | Integrated |
|--------------------------------------------|------|---------|--------------|------------|
| Accommodate water entry                    | 1998 | 0.21a   | 0.16b        | 0.23a      |
|                                            | 1999 | 0.21a   | 0.16b        | 0.20ab     |
| Facilitate water movement and availability | 1998 | 0.21a   | 0.21a        | 0.24b      |
|                                            | 1999 | 0.19a   | 0.18a        | 0.20a      |
| Resist surface structure degradation       | 1998 | 0.23ab  | 0.19a        | 0.24b      |
|                                            | 1999 | 0.21a   | 0.15b        | 0.21a      |
| Sustain fruit quality and productivity     | 1998 | 0.24a   | 0.23ab       | 0.21b      |
|                                            | 1999 | 0.22a   | 0.21a        | 0.21a      |
| Total soil quality rating                  | 1998 | 0.88a   | 0.78b        | 0.92a      |
|                                            | 1999 | 0.83a   | 0.70b        | 0.81a      |

Soil quality functions (each with a maximum value of 0.25) were assigned values on the basis of soil properties analysed and then added to determine soil quality ratings for each system. A total soil quality rating of 1.0 represents soil conditions optimal for both fruit production and environmental quality. Differences between values in a year followed by different letters are significant at the 0.05 level (least significant difference).





**Figure 1** Fruit yields of three apple production systems. Differences between values in a year followed by different letters are significant at the 0.05 level (least significant difference).

concept<sup>11</sup>, readily lend itself to quantification. To be sustainable, a farm must produce adequate yields of high quality, be profitable, protect the environment, conserve resources and be socially responsible in the long term<sup>12</sup>. But under conventional economic systems, market and social forces can change the viability of a production system independent of its environmental sustainability<sup>13</sup>. It has been proposed that ecological and economic systems should be linked so that ecosystem services are accounted for in commercial markets, thereby making sustainable land management a prerequisite for economic sustainability<sup>14,15</sup>.

A crucial step in developing such ecological-economic links is to assess the effects of agricultural systems on specific, measurable properties that are important indicators of sustainability. We measured the effects of an organic, a conventional and an integrated apple production system on the sustainability indicators of soil quality, horticultural performance, orchard profitability, environmental quality and energy efficiency. Perennial food crops such as apples may prove to be more sustainable to produce over the long term than annual crops<sup>16</sup>, and they currently comprise a significant portion of the world's agricultural production. For example, globally, nearly 5.6 million hectares of apples were harvested in 2000 (ref. 17). In the USA alone, apples and other high-value perennial food crops constituted 16% of the total value of food crops in 1998 (ref. 18).

We measured soil quality by analysing physical, chemical and biological soil properties and incorporating the data into a soil quality index<sup>19</sup>. Soil quality is the capacity of a soil to sustain

biological productivity, maintain environmental quality and promote plant and animal health<sup>20</sup>. We evaluated soil quality in terms of four soil functions: accommodating water entry; accommodating water movement and availability; resisting surface structure degradation; and supporting fruit quality and productivity. Soil quality ratings in 1998 and 1999 for the organic and integrated systems were significantly higher than those for the conventional system (Table 1), largely owing to the addition of compost and mulch in 1994 and 1995. Organic matter has a profound impact on soil quality, enhancing soil structure and fertility and increasing water infiltration and storage<sup>21</sup>. Because of poorer ability to accommodate water entry and to resist surface structure degradation, the conventional system (no organic amendments added) scored lowest overall in soil quality.

We assessed horticultural performance by measuring fruit yields, size and grade; tree growth; leaf and fruit mineral contents; fruit maturity; and consumer taste preference. There were no observable differences in pests, disease or physiological disorders among plots during each growing season. Differences in annual fruit yields were inconsistent among the three systems (Fig. 1). Cumulative yields were similar for all three systems. In 1995–1997, fruit size was similar across systems, except in 1996 when apples were larger in the integrated system (see Fig. A1 in Supplementary Information). In 1998 and 1999, the organic system produced smaller fruit (see Figs A1 and A2 in Supplementary Information). In 1995–1997, all marketed fruit produced from the three systems was sold for processing because it was downgraded primarily owing to skin russetting, a physiological skin disorder that reduces the fruit's visual appeal but not its taste or other attributes. (Although russeted Golden Delicious apples are not sold as fresh fruit in the US marketplace, Italy domestically markets a fully russeted Golden Delicious apple, and in the world market fully russeted Bosc pears are preferred to non-russeted ones.) The low landscape position of the experimental site in the orchard resulted in early season cool, humid conditions that contributed to the unusually high level of russetting. Fruit damage due to other physiological disorders, pests and diseases were minimal and equal for each of the three systems. In 1998 and 1999, marketable fruit not graded as Washington Extra Fancy or Fancy was sold for processing (see Table A1 in Supplementary Information).

Tree growth was similar in all three systems (see Fig. A3 in Supplementary Information). Although there were some differences in leaf nutrient contents among the three systems, analyses indicated satisfactory levels of nutrients<sup>22,23</sup> (see Table A2 in Supplementary Information). Fruit tissue nutrient analyses indicated

**Table 2** Gross receipts, total costs and net returns of apple production systems

| Year | Enterprise budget | Measured fruit quality     |                            |                            | Non-russeted fruit quality |                            |                            |
|------|-------------------|----------------------------|----------------------------|----------------------------|----------------------------|----------------------------|----------------------------|
|      |                   | Org (\$ ha <sup>-1</sup> ) | Con (\$ ha <sup>-1</sup> ) | Int (\$ ha <sup>-1</sup> ) | Org (\$ ha <sup>-1</sup> ) | Con (\$ ha <sup>-1</sup> ) | Int (\$ ha <sup>-1</sup> ) |
| 1994 | Gross receipts    | 0                          | 0                          | 0                          | 0                          | 0                          | 0                          |
|      | Total costs       | 23,486                     | 20,702                     | 23,740                     | 23,486                     | 20,702                     | 23,740                     |
|      | Net returns       | -23,486                    | -20,702                    | -23,740                    | -23,486                    | -20,702                    | -23,740                    |
| 1995 | Gross receipts    | 261a                       | 826b                       | 272a                       | 610a                       | 2,353b                     | 754a                       |
|      | Total costs       | 12,242a                    | 9,409b                     | 9,381b                     | 12,242a                    | 9,409b                     | 9,381b                     |
|      | Net returns       | -11,981a                   | -8,583b                    | -9,109c                    | -11,632a                   | -7,056b                    | -8,627c                    |
| 1996 | Gross receipts    | 2,482a                     | 4,559b                     | 3,294c                     | 5,825a                     | 13,001b                    | 9,149c                     |
|      | Total costs       | 11,608a                    | 11,917ab                   | 12,074b                    | 11,608a                    | 11,917ab                   | 12,074b                    |
|      | Net returns       | -9,126a                    | -7,358b                    | -8,780c                    | -5,784a                    | 1,084b                     | -2,925c                    |
| 1997 | Gross receipts    | 12,406a                    | 7,438b                     | 7,983b                     | 38,162a                    | 21,266b                    | 22,214b                    |
|      | Total costs       | 17,788a                    | 15,745b                    | 15,947b                    | 17,756a                    | 15,607b                    | 15,904b                    |
|      | Net returns       | -5,381a                    | -8,307b                    | -7,964b                    | 20,405a                    | 5,658b                     | 6,310b                     |
| 1998 | Gross receipts    | 16,485a                    | 8,967b                     | 9,211b                     | 37,442a                    | 21,691b                    | 21,640b                    |
|      | Total costs       | 19,074a                    | 17,863a                    | 17,906a                    | 19,637a                    | 17,682b                    | 18,143ab                   |
|      | Net returns       | -2,589a                    | -8,896b                    | -8,695b                    | 17,805a                    | 4,009b                     | 3,497b                     |
| 1999 | Gross receipts    | 24,503a                    | 21,846a                    | 21,438a                    | 25,821a                    | 23,653a                    | 22,593a                    |
|      | Total costs       | 18,714a                    | 20,188b                    | 20,011ab                   | 16,219a                    | 18,219b                    | 18,326b                    |
|      | Net returns       | 5,789a                     | 1,658a                     | 1,428a                     | 9,602a                     | 5,434a                     | 4,267a                     |

Measured from organic (Org), conventional (Con) and integrated (Int) apple production systems under measured fruit quality conditions (with excessive skin russetting) and non-russeted fruit quality conditions (with 15% cullage). Differences between values in a year and budget category followed by different letters are significant at the 0.05 level (least significant difference). No statistics were generated for 1994 data because costs were the same for each replicate.

some inconsistent differences (see Table A3 in Supplementary Information).

Mechanical analysis of fruit firmness at harvest and after storage in 1998 and 1999 showed that organic fruit was firmer (a positive consumer attribute for apples) than or as firm as conventional and integrated fruit (see Table A4 in Supplementary Information). Ratios of soluble solids (sugar) content to acidity (tartness), an indication of sweetness, were most often highest in organic fruit. These data were confirmed in taste tests by untrained sensory panels that found the organic apples to be sweeter after six months of storage than conventional apples and less tart at harvest and after six months storage than conventional and integrated apples (see Table A5 in Supplementary Information). The same taste tests, however, could not discern any difference in firmness among apples in the three systems at harvest or after storage. Taste tests also indicated that the integrated apples had a better flavour after six months storage but found no differences among organic, conventional and integrated apples in texture or overall acceptance.

Enterprise budgets were generated each year to calculate net returns from total costs and gross receipts. Receipts for the integrated system were estimated using prices for conventionally produced fruit, as unlike organic fruit there was no price premium for integrated fruit. Receipts for the organic system were estimated using prices for conventionally produced fruit in the first three years (1994–1996), the number of years necessary to convert from conventional to certified organic. The price premium to the grower for each grade of organic fruit in the next three years (1997–1999) averaged 50% above conventional prices.

The three systems did not show a net annual profit until 1999 under measured fruit quality conditions (with skin russetting) (Table 2). When we adjusted the economic analysis by eliminating the effects of russetting but maintained the estimated crop loss of 15% due to other factors and the measured size, grade and firmness of fresh fruit in this study, the organic system was more profitable than the conventional and integrated systems in 1997 and 1998. Higher production costs for the organic system in 1995 and 1997 (under measured and non-russetted fruit quality conditions) were largely due to differences in weed control practices, fruit thinning and compost applications. Production costs in 1999, however, were significantly lower for the organic system than for the other two systems due to reduced carryover interest costs resulting from faster repayment of the original investment.

The breakeven point, when cumulative net returns equal cumulative costs, can vary depending on several factors, such as fruit prices, input costs, yields and fruit quality. The breakeven point in

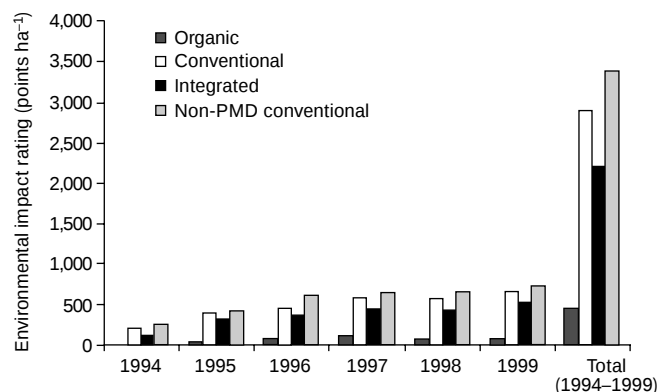
this study is projected to occur nine years after planting (in 2002) for the organic system under measured fruit quality conditions. The conventional and integrated systems would break even 15 and 17 years after planting, respectively, under measured conditions. Under non-russetted fruit quality conditions, the breakeven point would occur six, eight and nine years after planting for the organic, conventional and integrated systems, respectively. Assuming similar non-russetted fruit quality conditions, estimated breakeven points for conventional apple orchards in central Washington range from 8 to 11 years from planting<sup>24</sup>.

Without price premiums for organic fruit, the conventional system would break even first, the integrated second and the organic third under measured or non-russetted fruit quality conditions. For breakeven points of the organic and integrated systems to occur in the same year as the conventional system, price premiums of 12% for the organic system and 2% for the integrated system would be necessary under measured fruit quality conditions. Under non-russetted fruit quality conditions, premiums of 14% for the organic system and 6% for the integrated system would be necessary to match the breakeven point of the conventional system.

We assessed the environmental impact of the three production systems by using a rating index<sup>25</sup> employed by scientists and growers to determine the potential adverse impact of pesticides and fruit thinners, including naturally occurring certified organic products. The higher the rating, the greater the negative impact. As only 35% of conventional Washington apple growers use pheromone-mating disruption (PMD), an environmentally benign biological control used in our conventional treatment, we also included a conventional system in which synthetic pesticides were used in place of PMD. The total environmental impact rating of our conventional system was 6.2 times that of the organic system, whereas the integrated system rating was 4.7 times greater and the non-PMD conventional system rating was 7.7 times greater (Fig. 2).

Energy accounting was divided into inputs (labour, fuel, fertilizers and so on), output (yield) and output/input ratios (energy efficiency). Cumulative energy inputs and output for the six-year study period were lower for the organic system than for the conventional and integrated systems (Table 3). The output/input ratio for the organic system during the six-year study period, however, was 7% greater than that for the conventional system and 5% greater than that for the integrated system, making the organic system the most energy efficient.

Our results show that organic and integrated apple production systems in Washington State are not only better for soil and the environment than their conventional counterpart but have comparable yields and, for the organic system, higher profits and greater energy efficiency. Although crop yield and quality are important products of a farming system, the benefits of better soil and environmental quality provided by the organic and integrated



**Figure 2** Environmental impact ratings of four apple production systems: Organic, conventional, integrated and non-PMD conventional. Higher ratings indicate greater potential for negative environmental impact. For a listing of chemicals used and their impact ratings for the four production systems, see Table A6 in Supplementary Information.

**Table 3** Cumulative energy assessment

|                                       | Organic | Conventional | Integrated |
|---------------------------------------|---------|--------------|------------|
| Labour (h ha <sup>-1</sup> )          | 2,921   | 2,008        | 2,147      |
| Labour (MJ ha <sup>-1</sup> )         | 2,337   | 1,607        | 1,718      |
| Machinery (MJ ha <sup>-1</sup> )      | 73,974  | 73,560       | 73,560     |
| Fuel (MJ ha <sup>-1</sup> )           | 173,400 | 182,919      | 182,919    |
| Electricity (MJ ha <sup>-1</sup> )    | 10,794  | 10,794       | 10,794     |
| Fertilizer (MJ ha <sup>-1</sup> )     | 311*    | 16,255       | 8,901*     |
| Insecticide (MJ ha <sup>-1</sup> )    | 22,159  | 42,313       | 40,375     |
| Fungicide (MJ ha <sup>-1</sup> )      | 18,023  | 12,922       | 12,855     |
| Weed control (MJ ha <sup>-1</sup> )   | 141     | 31,931       | 13,350     |
| Infrastructure (MJ ha <sup>-1</sup> ) | 144,188 | 144,188      | 144,188    |
| Total input (MJ ha <sup>-1</sup> )    | 445,328 | 516,489      | 488,661    |
| Total output (MJ ha <sup>-1</sup> )   | 526,544 | 570,745      | 550,076    |
| Output/input (MJ MJ <sup>-1</sup> )   | 1.18    | 1.11         | 1.13       |

\*This includes composted poultry manure from a local commercial composting facility. The poultry facility from which the raw manure was obtained is assumed to be responsible for the energy charges required for composting the waste; consequently only energy requirements for transporting the compost to the orchard (63 MJ Mg<sup>-1</sup>) are charged against the organic and integrated systems<sup>30</sup>. For details on fertilizer inputs, see Table A7 in Supplementary Information.

production systems are equally valuable and usually overlooked in the marketplace. Such external benefits come at a financial cost to growers. Currently, growers of more sustainable systems may be unable to maintain profitable enterprises without economic incentives, such as price premiums or subsidies for organic and integrated products, that value these external benefits. Equally important, upon incorporation of external costs into economic assessments of farming systems, we may find that many currently profitable farming systems are uneconomical and therefore unsustainable. The challenge facing policymakers is to incorporate the value of ecosystem processes into the traditional marketplace, thereby supporting food producers in their attempts to employ both economically and environmentally sustainable practices. □

## Methods

### Study area

In May 1994, we planted four replicate plots for each of the three apple production systems with 'Golden Delicious' apples (*Malus × domestica* Borkh.) on EMLA.9 rootstocks on 1.7 ha in a randomized complete block design. The experiment was part of a 20-ha commercial apple orchard in the Yakima Valley, Washington. See Fig. A4 of Supplementary Information for plot layout and orchard system.

### Farming systems

In cooperation with the farmers, professional consultants and extension agents, we chose appropriate management practices for the three systems (see Table A7 in Supplementary Information). The organic system included compost and foliar sprays. In the first three years (1994–1996), bark mulch and landscape fabric controlled weeds; thereafter, cultivation and mowing were used for weed control. Organically certified biological controls, including applications of *Bacillus thuringiensis* and PMD to control codling moth (*Cydia pomonella* L.), were used for pest management. Fruit thinning was by hand. The conventional system included synthetic soil fertilizers and foliar sprays, pesticides, chemical fruit thinners and PMD. The integrated system used both compost and synthetic fertilizers and controlled weeds with both bark mulch and herbicides. Pest management and fruit thinning were similar to those of the conventional system. The three systems had similar total soil nitrogen inputs. Pests, diseases and physiological disorders were monitored throughout each growing season by the farmers and professional consultants, who recommended organic, conventional or integrated treatments for their control.

### Soil analyses

All soil samples were taken from the inner two rows of each experimental plot to minimize edge effects, excluding the first 20 trees from each end of these sample rows. Samples were collected midway between trees within tree rows. In 1998 and 1999, soil analyses included bulk density, water content, total nitrogen, nitrate-nitrogen, extractable phosphorus, cation-exchange capacity, pH, electrical conductivity, organic carbon content, aggregate stability, microbial biomass carbon and nitrogen, and earthworm populations. Details of analytical procedures are described elsewhere<sup>17</sup>.

### Horticultural performance

All components of horticultural performance were measured from trees, leaves and fruit sampled in the middle two rows of each plot, excluding the first 20 trees from each end of these sample rows. We recorded yields and size (average mass) of fruit at harvest in 1995–1999. The proportions of fruit suitable only for processing (due to small size or defects) and fruit suitable for fresh market and divided by grade were also recorded. Grading of fresh fruit was based on Washington State's apple industry standards<sup>26</sup>. We used trunk cross-sectional area to estimate unit tree growth. We analysed leaf mineral contents (N, P, K, S, Ca, Mg, B, Zn, Mn, Cu and Fe) in 1994–1999 from pooled samples of mid-shoot leaves taken randomly from each plot in midsummer. We analysed fruit flesh mineral contents (N, P, K, Ca, Mg, B and Zn) in 1995 and 1997–1999 from pooled samples of uniformly sized fruits taken three weeks before harvest. Mineral analyses were carried out according to standard methods<sup>23</sup>. We assessed fruit maturity parameters, including flesh firmness, soluble solids and acidity, according to standard procedures<sup>27</sup> at harvest and after three and six months of controlled atmosphere storage in 1998 and 1999. Untrained sensory panels were used to determine preferences for overall acceptance, texture, flavour, firmness, sweetness and tartness of 1999 fruit from each production system and storage treatment.

### Economic analyses

We calculated gross receipts using farmgate prices paid by packing houses to farmers for apples sold at harvest or after storage. Prices for the specific size, grade and firmness of 'Golden Delicious' organic and conventional apples from our study were based on prices from *Washington Growers Clearing House Bulletins* and fruit packing houses in Washington State. Total costs included non-harvest variable costs (fertilizers, pesticides, fuel, labour and water), harvest variable costs (picking, grading, packing and storage) and fixed costs (machinery, interest and taxes). Projected returns for 2000 and beyond were

estimated from average fruit sizes (1998–1999) and yields (1997–1999) for each treatment, assuming a 15% cullage rate for all three treatments and a 50% price premium for organic fruit.

### Environmental impact assessment

We determined environmental impact ratings for each farming system using an index developed by Stemilt Growers, Inc. of Wenatchee, Washington, as part of their 'Responsible Choice' program<sup>25</sup>. Similar to Cornell University's Environmental Impact Quotient<sup>28</sup> but updated to include fruit thinners and certified organic products, the index takes into account chemical efficacy, potential worker and consumer exposure, leaching potential, soil sorption index, chemical half-life and the effects of chemicals on beneficial organisms, all based on toxicological studies and chemical characteristics of each product. The active ingredient of each pesticide and the dose and frequency of application were used to calculate the environmental impact ratings.

### Energy use

Energy use for each production system was calculated from energy data specific to agricultural production<sup>29,30</sup>.

Received 19 January; accepted 15 March 2001.

- Matson, P. A., Parton, W. J., Power, A. G. & Swift, M. J. Agricultural intensification and ecosystem properties. *Science* **277**, 504–509 (1997).
- Drinkwater, L. E., Wagoner, P. & Sarantonio, M. Legume-based cropping systems have reduced carbon and nitrogen losses. *Nature* **396**, 262–265 (1998).
- Tillman, D. Global environmental impacts of agricultural expansion: the need for sustainable and efficient practices. *Proc. Natl Acad. Sci. USA* **96**, 5995–6000 (1999).
- Zhu, Y. *et al.* Genetic diversity and disease control in rice. *Nature* **406**, 718–722 (2000).
- Avery, D. in *Fearing Food: Risk, Health and Environment* (eds Morris, J. & Bate, R.) 3–18 (Butterworth-Heinemann, Oxford, 1999).
- Trewavas, A. Much food, many problems. *Nature* **402**, 231–232 (1999).
- Greene, C. U. S. Organic agriculture gaining ground. *Agric. Outlook* **270**, 9–14 (2000).
- Hamm, U. & Michelsen, J. in *IFOAM 2000—The World Grows Organic. Proc. 13th Int. IFOAM Scientific Conf.* (eds Alföldi, T., Lockeretz, W. & Niggli, U.) 507–510 (IOS, Zurich, 2000).
- Sansavini, S. Integrated fruit production in Europe: Research and strategies for a sustainable industry. *Sci. Hort.* **68**, 25–36 (1997).
- Reganold, J. Farming's organic future. *New Sci.* **122**, 49–52 (1989).
- Barrett, G. W. & Odum, E. P. The twenty-first century: The world at carrying capacity. *BioScience* **50**, 363–368 (2000).
- Reganold, J. P., Papendick, R. I. & Parr, J. F. Sustainable agriculture. *Sci. Am.* **262**, 112–120 (1990).
- Goodland, R. & Daly, H. Environmental sustainability: Universal and nonnegotiable. *Ecol. Appl.* **6**, 1002–1017 (1997).
- Costanza, R. *et al.* The values of the world's ecosystem services and natural capital. *Nature* **387**, 253–260 (1997).
- Naveh, Z. The total human ecosystem: Integrating ecology and economics. *BioScience* **50**, 357–361 (2000).
- Merwin, I. A. & Pritts, M. P. Are modern fruit production systems sustainable? *HortTechnology* **3**, 128–136 (1993).
- FAO Crop Production Statistics Web site [online], (cited 3 April 2001) (<http://apps.fao.org/lpage/collections?subset=agriculture>) (2001).
- USDA National Agricultural Statistics Service *Agricultural Statistics 1999 IX–17–IX–19* (US Government Printing Office, Washington DC, 1999).
- Glover, J. D., Reganold, J. P. & Andrews, P. K. Systematic method for rating soil quality of conventional, organic, and integrated apple orchards in Washington State. *Agric. Ecosystem Env.* **80**, 29–45 (2000).
- Doran, J. W. & Parkin, T. B. in *Defining Soil Quality for a Sustainable Environment* (eds Doran, J. W., Coleman, D. C., Bezdicek, D. F. & Stewart, B. A.) 3–21 (Soil Sci. Soc. Am., Madison, Wisconsin, 1994).
- Brady, N. C. & Weil, R. R. *The Nature and Properties of Soils* 12th edn 468–469 (Prentice Hall, Upper Saddle River, New Jersey, 1999).
- Dow, A. I. *Critical Nutrient Ranges for Northwest Crops* (West. Reg. Ext. Pub. 43, Oregon State University Cooperative Extension Service, Corvallis, 1980).
- Gavlak, R. G., Horneck, D. A. & Miller, R. O. *Plant, Soil and Water Reference Methods for the Western Region* (West. Reg. Ext. Pub. 125, Univ. Alaska Cooperative Extension Service, Fairbanks, 1994).
- Bechtel, L., Barritt, B., Dilley, M. & Hinman, H. *Economic Analysis of Apple Orchard Management Systems with Three Varieties in Central Washington* (Washington State Univ. Research Bulletin XB1032, Pullman, 1995).
- Reed, N. in *125th Annual Report of the Secretary of the State Horticultural Society of Michigan* (ed. Hull, J. Jr) 68–78 (Mich. State Hort. Soc., East Lansing, 1995).
- Revised Code of Washington (RCW). Title 15. Chapter 15. 17. *Standards of Grades and Packs* (West Publishing, St Paul, 1998).
- Apple Maturity Program Handbook* (Apple Maturity Program, Wenatchee, Washington, 1986).
- Kovach, J., Petzoldt, C., Degni, J. & Tette, J. A *Method to Measure the Environmental Impact of Pesticides* (New York's Food and Life Sciences Bulletin No. 139, Cornell Univ., Ithaca, 1992).
- Leach, G. *Energy and Food Production* 71–77 (IPC Science and Technology Press, Guildford, UK, 1976).
- Funt, R. C. in *Handbook of Energy Utilization in Agriculture* (ed. Pimentel, D.) 235–246 (CRC, Boca Raton, Florida, 1980).

**Supplementary information** is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.



## Acknowledgements

We thank A. and E. Dolph for the use of their farm. We acknowledge funding from the USDA Agricultural Systems Program. We thank D. Fahy, D. Granatstein, L. Klein, S. Leach, C. Litzinger, A. O'Rourke, A. Palmer, R. Papendick, P. Salant, S. Sansavini and T. Schotzko for comments on drafts of this manuscript, and R. Alldredge, S. Drake, M. Fauci, J. Fellman, S. Mattinson, J. Powers, N. Reed, R. Rupp and K. Weller for technical assistance.

Correspondence and request for materials should be addressed to J.P.R. (e-mail: reganold@wsu.edu).

## The concepts of 'sameness' and 'difference' in an insect

Martin Giurfa<sup>\*,†</sup>, Shaowu Zhang<sup>‡</sup>, Arnim Jenett<sup>\*</sup>, Randolph Menzel<sup>\*</sup> & Mandyam V. Srinivasan<sup>‡</sup>

<sup>\*</sup> Neurobiologie, Institut für Biologie, Freie Universität Berlin, Königin-Luise-Strasse 28/30, 14195 Berlin, Germany

<sup>†</sup> Ethologie et Cognition Animale, Université Paul Sabatier, 118 Route de Narbonne, 30162 Toulouse, Cedex 4, France

<sup>‡</sup> Centre for Visual Science, Research School of Biological Sciences, Australian National University, PO Box 475, Canberra, ACT 2601, Australia

Insects process and learn information flexibly to adapt to their environment. The honeybee *Apis mellifera* constitutes a traditional model for studying learning and memory at behavioural, cellular and molecular levels<sup>1</sup>. Earlier studies focused on elementary associative and non-associative forms of learning determined by either olfactory conditioning of the proboscis extension reflex<sup>1</sup> or the learning of visual stimuli<sup>2</sup> in an operant context. However, research has indicated that bees are capable of cognitive performances that were thought to occur only in some vertebrate species. For example, honeybees can interpolate visual information<sup>3</sup>, exhibit associative recall<sup>4,5</sup>, categorize visual information<sup>6–8</sup> and learn contextual information<sup>9</sup>. Here we show that honeybees can form 'sameness' and 'difference' concepts. They learn to solve 'delayed matching-to-sample' tasks, in which they are required to respond to a matching stimulus, and 'delayed non-matching-to-sample' tasks, in which they are required to respond to a different stimulus; they can also transfer the learned rules to new stimuli of the same or a different sensory modality. Thus, not only can bees learn specific objects and their physical parameters, but they can also master abstract inter-relationships, such as sameness and difference.

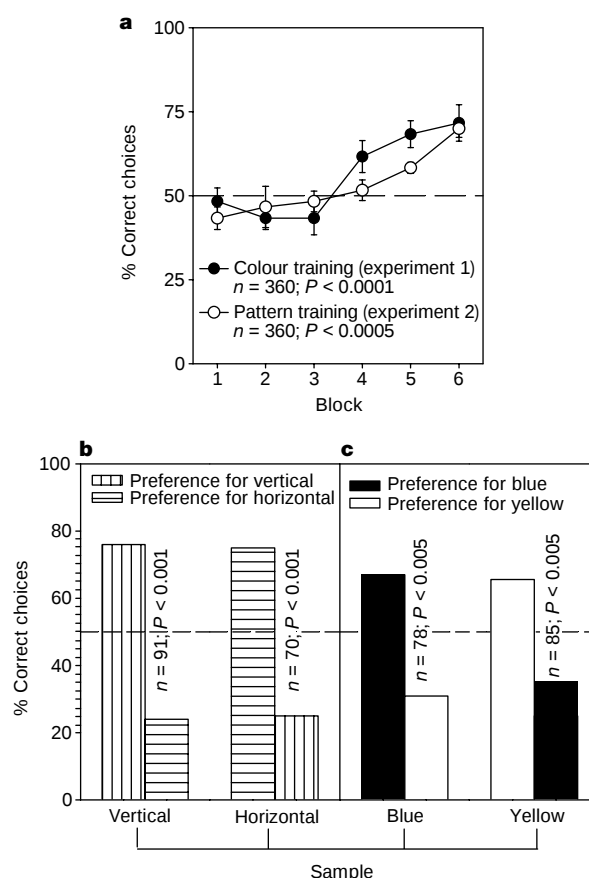
An important cognitive capacity is the ability to learn relationships between stimuli. In vertebrates, the capacity to acquire sameness–difference concepts has been studied using two experimental procedures, the matching task and the oddity task. A variation of the former is the 'delayed matching-to-sample' task, in which an animal is presented with a 'sample' and subsequently with two or more secondary stimuli, one of which is identical to the sample. The animal is required to respond to the stimulus just encountered. The 'delayed non-matching-to-sample' task is a variation of the oddity task. The procedure is similar to the matching-to-sample task except that the animal is required to respond to the stimulus that is different from the sample. In both cases, broadly construed sameness and difference concepts are shown only if the animal exhibits positive transfer to a completely new set of stimuli, which it had not experienced during training.

We trained honeybees, *A. mellifera*, in a delayed matching-to-sample paradigm to examine whether they could form a concept of sameness. Training was carried out using a Y-maze placed close to a

laboratory window. Each bee entered the maze by flying through a hole in the middle of an entrance wall. At the entrance, the bee encountered the sample stimulus. The sample was one of two different stimuli, A or B, alternated in a pseudo-random sequence. The entrance led to a decision chamber, where the bee could choose one of two arms. Each arm carried either stimulus A or stimulus B as secondary stimulus. The bee was rewarded with sucrose solution only if it chose the stimulus that was identical to the sample. If the bees managed to learn the original discrimination, they were presented with new sample and secondary stimuli in 'transfer tests': the bees had to choose between stimuli C and D, when the sample was either C or D.

Four experiments were carried out, each by training a fresh group of bees. In experiment 1, A and B were colours, whereas C and D were vertical and horizontal black and white gratings. In experiment 2, A and B were gratings, whereas C and D were colours. In experiment 3, A and B were radial and circular gratings, whereas C and D were oriented linear gratings. In experiment 4, A and B were odours, whereas C and D were colours. In all experiments the stimuli chosen were well distinguished by the bees, as indicated by preliminary investigations.

The results for experiments 1 and 2 are shown in Fig. 1. In experiment 1, the bees were trained to match a given colour (blue versus yellow). The acquisition curve showed a significant improvement during training (Fig. 1a:  $P < 0.0001$ ): bees preferred the colour that was identical to the sample, independently of the sample



**Figure 1** Experiments 1 and 2. **a**, Learning performance of bees during training on colours (experiment 1) and on vertical and horizontal gratings (experiment 2). Data show the results of blocks of ten consecutive training visits for each experiment. **b**, **c**, Results of transfer tests. **b**, Transfer tests with patterns (colour training). In experiment 1, bees trained on the colours were tested on the gratings. **c**, Transfer tests with colours (pattern training). In experiment 2, bees trained on the gratings were tested on the colours. n, number of choices evaluated.

colour. In experiment 2, the bees were trained to match a given grating orientation (vertical versus horizontal). The acquisition curve also increased significantly along the training (Fig. 1a;  $P < 0.0005$ ): bees preferred the orientation of the grating that was identical to that of the sample, independently of the sample orientation. At the end of the learning phase, the choice frequency for the matching stimulus was above 70% in both experiments 1 and 2.

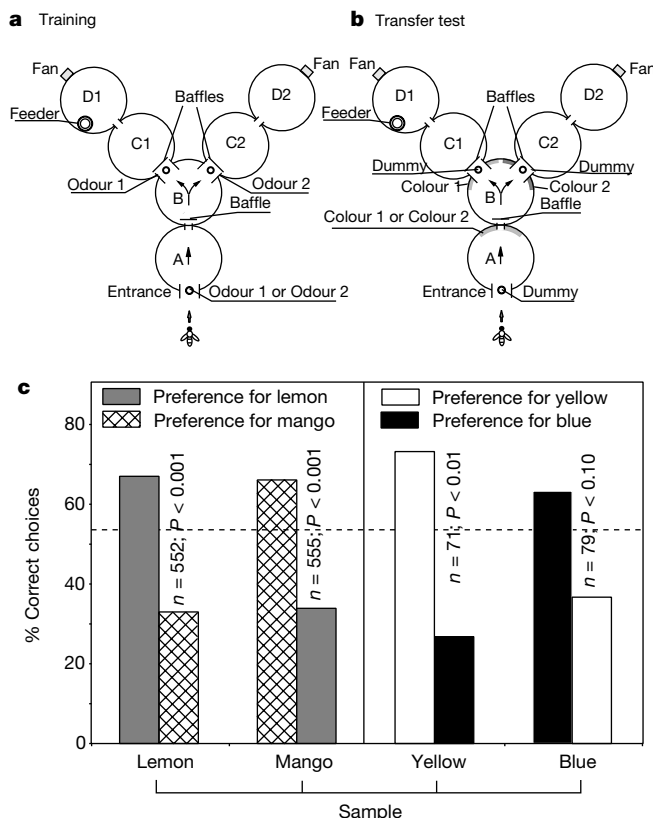
These results, however, do not necessarily show a capacity to learn the concept of sameness between stimuli. Bees could solve the problem by simply learning the appropriate choice for each of the configurations used in training (for example, in experiment 1 'choose blue and not yellow if blue was seen at the entrance' and 'choose yellow and not blue if yellow was seen at the entrance'), without abstracting a general concept of sameness. The capacity to abstract a general concept of sameness is shown in the transfer tests, where the bees were confronted with new samples and secondary stimuli. In such tests, bees that had been trained to match the colours (experiment 1) could match the gratings (Fig. 1b); and bees that had been trained to match the gratings (experiment 2) could match the colours (Fig. 1c).

The performance in these transfer tests was similar to that attained in the training. In the transfer tests of experiment 1 (Fig. 1b), the matching orientation was chosen with a frequency of 76.3% ( $n = 91$ ,  $P < 0.001$ ) if the sample was a vertical grating, and 75.0% ( $n = 70$ ,  $P < 0.001$ ) if the sample was a horizontal grating. The difference between the preferences for vertical in the two cases was highly significant ( $P < 0.001$ ). In the transfer tests of experiment 2 (Fig. 1c), the matching colour was chosen with a frequency of 67.3% ( $n = 78$ ,  $P < 0.005$ ) if the sample was blue, and 65.3%

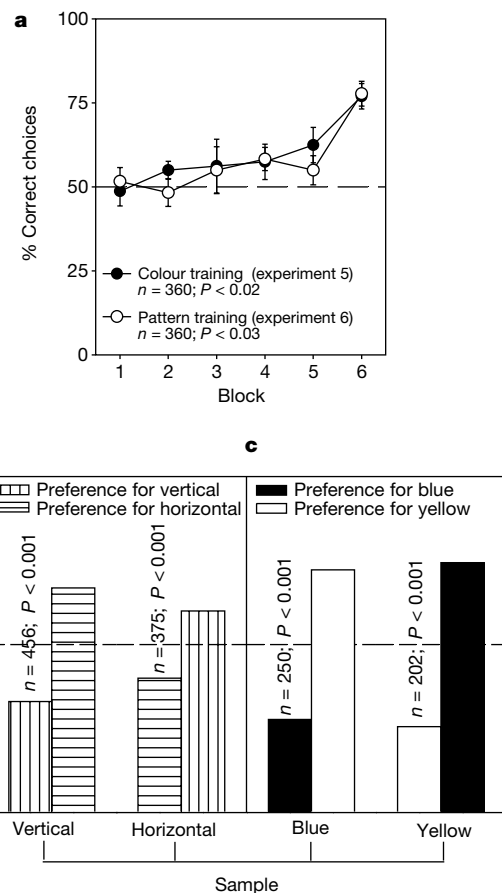
( $n = 85$ ,  $P < 0.005$ ) if the sample was yellow. The difference between the preferences for blue in the two cases was highly significant ( $P < 0.001$ ).

Similar results were obtained in experiment 3, in which the bees were trained with radial and circular gratings. After acquisition, the bees chose the matching stimulus with a frequency of 78.0% ( $n = 477$ ,  $P < 0.001$ ) if the sample was a radial grating, and of 77.7% ( $n = 491$ ,  $P < 0.001$ ) if it was a circular grating. In the transfer tests using linear gratings (oriented at  $+45^\circ$  and  $-45^\circ$ ), the trained bees chose the appropriate grating with a frequency of 59.8% ( $n = 92$ ,  $0.1 < P < 0.2$ ) if the sample was a  $45^\circ$  grating, and of 63.9% ( $n = 119$ ,  $P < 0.05$ ) if it was a  $-45^\circ$  grating. The difference between the preferences for the  $45^\circ$  grating in the two cases was highly significant ( $P < 0.001$ ). As in experiment 2, bees that had been trained on the radial and circular gratings could transfer their learned matching ability to colours (data not shown).

In experiment 4, we explored matching in a different stimulus modality, as well as transfer across stimulus modalities. The bees were trained to match odours (lemon and mango; Fig. 2). After acquisition, the trained bees showed a significant preference for the matching odour (Fig. 2c, left). When the trained bees were subjected to transfer tests using colours (blue and yellow), they showed a preference for the appropriate colour (Fig. 2c, right). Details and statistics are given in Fig. 2. The difference between the preferences for yellow in the two cases was highly significant ( $P < 0.001$ ). Thus,



**Figure 2** Experiment 4. **a**, Set-up in which bees were trained to match odours. **b**, Set-up used for transfer tests in which bees trained on odour matching were tested on colour matching. **c**, Results. Left, learning tests for odour matching; right, transfer tests for colour matching.



**Figure 3** Experiments 5 and 6, investigating non-matching performance. **a**, Learning performance of bees during training on colours (experiment 5) and on vertical and horizontal gratings (experiment 6). Data show the results of blocks of ten consecutive training visits for each experiment. **b**, **c**, Results of transfer tests. **b**, Transfer tests with patterns (colour training). In experiment 5, bees trained on the colours were tested on the gratings. **c**, Transfer tests with colours (pattern training). In experiment 6, bees trained on the gratings were tested on the colours.  $n$ , number of choices evaluated.

bees can learn to perform a delayed match-to-sample task even in the odour domain. Furthermore, they can transfer this capacity to the visual domain.

Finally, we trained bees in a delayed non-matching-to-sample task to examine whether they could form a concept of difference. Training was carried out as in experiments 1 and 2, except that in both cases bees were rewarded only if they chose the stimulus that was different from the sample. In experiment 5, bees were trained with the colours yellow and blue and tested with the vertical and horizontal black and white gratings. In experiment 6, bees were trained with the gratings and tested with the colours.

In experiment 5, the bees learned to choose the colour that was different from that of the sample. The acquisition curve showed a significant improvement during training (Fig. 3a;  $P < 0.02$ ). In experiment 6, the bees learned to choose the grating orientation that was different from that of the sample. The acquisition curve also increased significantly along the training (Fig. 3a;  $P < 0.03$ ). At the end of the training, the choice frequency for the matching stimulus was above 70% in both experiments.

In the transfer tests, the bees that had been trained to choose the different colour (experiment 5) transferred their choice to the different grating (Fig. 3b); and bees that had been trained to choose the different grating (experiment 6) transferred their choice to the different colour (Fig. 3c). In all cases, the bees showed a significant preference for the non-matching sample (for details and statistics, see Fig. 3).

Animals can learn matching or oddity tasks by using simple rules such as similarity judgements if they are trained and tested on the same stimuli. Our experiments reveal, however, that bees can do more than this: they can transfer the matching or non-matching ability to new stimuli for which the rules of choice are not specified in the training. Because the bees continue to choose the appropriate matching (or non-matching) stimulus even in new situations, we conclude that they can form and use a concept of sameness (experiments 1 to 4) and difference (experiments 5 and 6) in making their choices. The ability to judge sameness and difference has been shown in many vertebrates<sup>10–20</sup>, but in some cases the results seem to depend critically on procedural details<sup>20</sup>. Furthermore, transfer from one problem to another, when shown, is usually very weak<sup>14</sup>.

Our findings with bees, however, are very consistent. They were repeatable in two different laboratories that used different experimental set-ups, test procedures and colonies. The concepts of sameness and difference, once learned, can be transferred to other stimuli that belong either to the same domain (as in the transfer from radial and circular gratings to oriented linear gratings) or to a different domain (as in the transfer from colours to oriented linear gratings and vice versa). Furthermore, the concept of sameness can be learned even in the olfactory modality, and can be transferred across stimulus modalities (from odours to colours). Thus, in the bee, the ability to form 'sameness-difference' concepts is broad and robust. Such concepts might contribute to improve foraging activities.

Our results question the view that vertebrates, and in particular primates, may be the only animals able to form 'sameness' or 'oddity' concepts<sup>21</sup>. They also show that higher cognitive functions are not a privilege of vertebrates<sup>22</sup>. Moreover, as such functions result from a relatively simple and accessible nervous system in which elementary associative learning pathways have been already characterized<sup>23</sup>, there is a realistic chance of uncovering the neural mechanisms that underlie this capacity. □

## Methods

### Bees and set ups

For each experiment, a group of 6–8 individually marked bees were trained to enter and collect 50% (weight/weight) sucrose solution in a Y-maze placed close to an open window in the laboratory. Experiments 1, 2, 5 and 6 were conducted in Berlin; experiments 3 and 4 were conducted in Canberra. All mazes were covered by an ultraviolet-transmitting

Plexiglas ceiling. Bees entered the maze by flying through an aperture, 4 cm in diameter.

The set-up for training the olfactory stimuli is shown in Fig. 2a. Bees received odour 1 as they entered chamber A. They then entered chamber B, where they were presented with two odours (1 and 2) and had to learn to choose the odour matching that at the entrance. The odours were presented by perforated vials containing tissue paper soaked in liquid odorant. To prevent mixing of odours in chamber B, and to remove any feeder odours in the reward chambers D1 and D2, suction fans were run for 5–10 s immediately after each time the stimulus positions in the decision chamber were interchanged; that is, every 10 min on average. Bees trained with odours in this way were subsequently tested for colour matching, using the apparatus shown in Fig. 2b. Here the odours were replaced by colours, and the odour vials were replaced by dummy vials which carried no scent, but had the same visual appearance as the vials used in the training. Baffles prevented the bees from experiencing the stimuli in chamber B until they had entered it, and from viewing the feeder from chamber B.

### Visual stimuli

The visual stimuli used in the experiments were 11-cm diameter discs presented in the vertical plane. The colour stimuli were blue and yellow, cut from HKS-N 44 and 3 cardboards, respectively (K+E Stuttgart). All gratings were black and white and were printed using a high-resolution laser printer. The linear gratings had a period of 3 cm, the annular gratings had a period of 4 cm, and the radial gratings had an angular period of 60°.

### Training and experiments

During training in all experiments, the position (left or right) of the rewarded stimulus in the maze was changed after every two visits per bee (on average), and the sample was changed after every four visits per bee. We evaluated the learning of the training discrimination by determining significant changes in the acquisition curves (experiments 1, 2, 5 and 6) or by learning tests in which the performance was quantified by recording the choices of each bee after it entered the maze (experiments 3 and 4). In the latter case choices were registered as the first entry by the flying bee through an aperture that was associated with each stimulus.

Transfer tests were conducted after 60 training visits per bee (equivalent to one day of training) in experiments 1, 2, 3, 5 and 6 (which involved training on visual stimuli) and after 80 visits in experiment 4 (which involved training on olfactory stimuli). No reward was offered during these tests. Performance was recorded either as touches (flights towards a stimulus that ended with the antennae touching the stimulus surface) during 2 min (experiments 1, 2, 5 and 6) or as in the learning tests (experiments 3 and 4). In all cases, each test was divided into four parts: two with one stimulus as the sample and two with the alternative stimulus as the sample. Tests with a given sample were performed twice: once with the matching stimulus presented in the right arm of the maze, and once in the left arm. Between transfer tests, the bees were retrained (usually 5–10 visits to the apparatus) to maintain the level of learning at its plateau.

As the stimuli in the transfer tests were different from those used in the training, bees were sometimes reluctant to enter the apparatus and make a choice during the transfer tests. To eliminate this difficulty, in experiments 3 and 4 a 'familiarization procedure' was used from time to time during the training, in which the set-up presented the sample and secondary stimuli that would eventually be used in the transfer test. However, the feeder was placed at a 'neutral' position beyond the entrance, at the point where they would normally choose between the stimuli in the two arms. This procedure ensured that the bees became accustomed to the transfer stimuli, without being trained to choose between them.

### Statistical analysis

We used analysis of variance for repeated measurements to determine whether the bees' performance in experiments 1, 2, 5 and 6 significantly improved as a result of the training. In the learning tests in experiments 3 and 4 and in the transfer tests in all experiments,  $\chi^2$  tests were used to determine whether the experimentally measured choice frequencies were significantly different from random-choice. We used  $\chi^2 \times 2$  tests to determine whether two choice frequencies were significantly different from each other.

Received 19 October 2000; accepted 29 January 2001.

1. Menzel, R. & Müller, U. Learning and memory in honeybees: from behavior to neural substrates. *Annu. Rev. Neurosci.* **19**, 379–404 (1996).
2. Menzel, R. in *Experimental Behavioral Ecology* (eds Hölldobler, B. & Lindauer, M.) 55–74 (Gustav Fischer, Stuttgart, 1985).
3. Collett, T. S. & Baron, J. Learnt sensori-motor mappings in honeybees: interpolation and its possible relevance to navigation. *J. Comp. Physiol. A* **177**, 287–298 (1995).
4. Srinivasan, M. V., Zhang, S. W. & Zhu, H. Honeybees links sights to smells. *Nature* **396**, 637–638 (1998).
5. Zhang, S. W., Lehrer, M. & Srinivasan, M. V. Honeybee memory: navigation by associative grouping and recall of visual stimuli. *Neurobiol. Learn. Mem.* **72**, 180–201 (1999).
6. Hateren, J. H., Srinivasan, M. V. & Wait, P. B. Pattern recognition in bees: orientation discrimination. *J. Comp. Physiol. A* **167**, 649–654 (1990).
7. Giurfa, M., Eichmann, B. & Menzel, R. Symmetry perception in an insect. *Nature* **382**, 458–461 (1996).
8. Horridge, G. A. & Zhang, S. W. Pattern vision in honeybees (*Apis mellifera*): flower-like patterns with no predominant orientation. *J. Insect Physiol.* **41**, 681–688 (1994).
9. Collett, T. S., Fauria, K., Dale, K. & Baron, J. Places and patterns—a study of context learning in honeybees. *J. Comp. Physiol. A* **181**, 343–353 (1997).
10. Moon, L. E. & Harlow, H. F. Analysis of oddity learning by rhesus monkeys. *J. Comp. Physiol. Psychol.* **48**, 188–194 (1955).
11. Strong, P. N. & Hedges, M. Comparative studies in simple oddity learning. 1. Cats, racoons, monkeys and chimpanzees. *Psychonom. Sci.* **5**, 13–14 (1966).



12. Thomas, R. K. & Boyd, M. G. A comparison of *Cebus albifrons* and *Samiri sciureus* on oddity performance. *Anim. Learn. Behav.* **1**, 151–153 (1973).
13. Herman, L. M. & Gordon, J. A. Auditory delayed matching in the bottlenose dolphin. *J. Exp. Anal. Behav.* **21**, 19–26 (1974).
14. Zentall, T. R. & Hogan, D. E. Same/different concept learning in the pigeon: the effect of negative instances and prior adaptation to transfer stimuli. *J. Exp. Anal. Behav.* **30**, 177–186 (1978).
15. Holmes, P. W. Transfer of matching performance in pigeons. *J. Exp. Anal. Behav.* **31**, 103–114 (1979).
16. Lombardi, C. M., Fachinelli, C. G. & Delius, J. D. Oddity of visual patterns conceptualized by pigeons. *Anim. Learn. Behav.* **12**, 2–6 (1984).
17. D'Amato, M. R., Salmon, D. P. & Colombo, M. Extent and limits of the matching concept in monkeys (*Cebus apella*). *J. Exp. Psychol. Anim. Behav. Proc.* **11**, 35–51 (1985).
18. Wilson, B., Mackintosh, N. J. & Boakes, R. A. Transfer of relational rules in matching and oddity learning by pigeons and corvids. *Quart. J. Exp. Psychol.* **37B**, 313–332 (1985).
19. Oden, D. L., Premack, D. & Thompson, R. K. R. Spontaneous transfer of matching by infant chimpanzees. *J. Exp. Psychol. Anim. Behav. Proc.* **14**, 140–145 (1988).
20. Wright, A. A., Santiago, H. C., Urciuoli, P. J. & Sands, S. F. in *Quantitative Analyses Of Behavior: Discrimination Processes* Vol. IV (eds Commons, M. L., Herrnstein, R. J. & Wagner, A. R.) 295–317 (Ballinger, Cambridge MA, 1984).
21. Davey, G. *Ecological Learning Theory* (Routledge, London, 1989).
22. Menzel, R. & Giurfa, M. Cognitive architecture of a mini-brain: the honeybee. *Trends Cognit. Sci.* **5**, 62–71 (2001).
23. Hammer, M. The neural basis of associative reward learning in honeybees. *Trends Neurosci.* **20**, 245–252 (1997).

#### Acknowledgements

We thank C. Bönisch and C. Wiley for help with experiments 5 and 6. M.G., A.J. and R.M. were supported by the Special Programme SFB 515 of the Deutsche Forschungsgemeinschaft. M.G. was also supported by a grant from the Deutsche Forschungsgemeinschaft. S.W.Z. and M.S. were supported partly by a grant from the Human Frontiers in Science Program and from the US Defence Advanced Research Projects Agency and the Office of Naval Research. The two laboratories contributed equally to the work.

Correspondence and requests for materials should be addressed to M.G. (e-mail: giurfa@cict.fr) or S.W.Z. (e-mail: swzhang@rsbs.anu.edu.au).

## Chromatic sensitivity of ganglion cells in the peripheral primate retina

Paul R. Martin\*, Barry B. Lee†‡, Andrew J. R. White\*, Samuel G. Solomon\* & Lukas Rüttiger†

\* Department of Physiology and Institute for Biomedical Research, University of Sydney, NSW 2006, Australia

† Neurobiology group, Max-Planck-Institute for Biophysical Chemistry, D37077 Göttingen, Germany

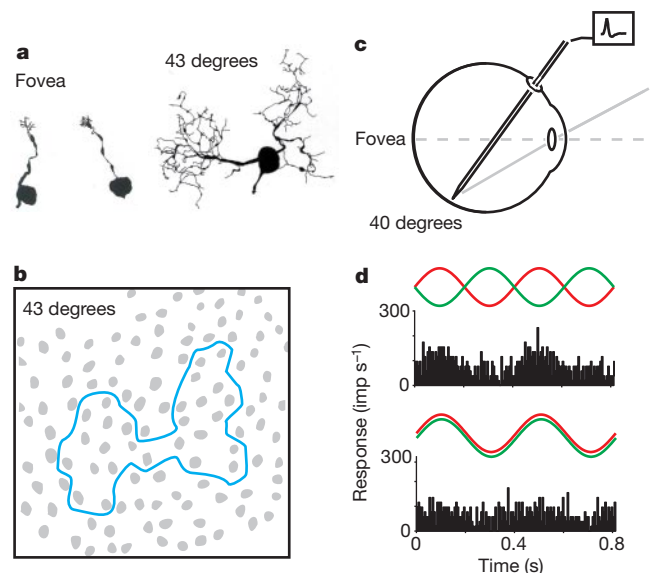
‡ State College of Optometry, SUNY, New York, New York 10036-8003, USA

Visual abilities change over the visual field. For example, our ability to detect movement is better in peripheral vision than in foveal vision, but colour discrimination is markedly worse<sup>1,2</sup>. The deterioration of colour vision has been attributed to reduced colour specificity in cells of the midget, parvocellular (PC) visual pathway in the peripheral retina<sup>3–5</sup>. We have measured the colour specificity (red–green chromatic modulation sensitivity) of PC cells at eccentricities between 20 and 50 degrees in the macaque retina. Here we show that most peripheral PC cells have red–green modulation sensitivity close to that of foveal PC cells. This result is incompatible with the view that PC pathway cells in peripheral retina make indiscriminate connections ('random wiring') with retinal circuits devoted to different spectral types of cone photoreceptors<sup>4,6,7</sup>. We show that selective cone connections can be maintained by dendritic field anisotropy, consistent with the morphology of PC cell dendritic fields in peripheral retina<sup>8,9</sup>. Our results also imply that postretinal mechanisms contribute to the psychophysically demonstrated deterioration of colour discrimination in the peripheral visual field.

In primates, ganglion cells of the midget, parvocellular pathway (PC cells) are proposed to be the origin of the red–green channel for

colour vision<sup>3,10–12</sup>. Most foveal PC cells show chromatic opponent responses: they are excited by some wavelengths in the visible spectrum and inhibited by others. Chromatic opponency of PC cells in the fovea is thought to depend on the fact that they derive input to the receptive field centre from an individual middle (M) or long (L) wavelength-sensitive cone through a midget bipolar cell<sup>4,6,13</sup>. In peripheral retina, the receptive field centres (and dendritic fields) of PC cells are much larger than those in the fovea, and encompass the area of 20–40 M and L cones<sup>8,9</sup> (Fig. 1a, b). The M and L cones show little or no spatial clustering<sup>14,15</sup>; therefore, if the peripheral PC cells draw their receptive field centre indiscriminately from the cones in their dendritic field, they will receive mixed spectral input, and show reduced red–green chromatic sensitivity in comparison to their foveal counterparts. Mixed spectral input to PC cells has been proposed as the basis for the decline in chromatic sensitivity in peripheral vision (the 'random wiring' hypothesis)<sup>3–5</sup>.

To test directly the random wiring hypothesis, we made extra-cellular single-cell recordings from the intact eye of anaesthetized macaque monkeys using standard techniques<sup>12,16</sup>. Recording sites were concentrated in peripheral retina between 20 and 50 degrees eccentricity (Fig. 1c). Cells were classified by receptive field tests<sup>11,17,18</sup>, including measurement of receptive field dimensions, and measurement of contrast sensitivity using luminance-matched red and green light-emitting diodes (LEDs; Fig. 1d)<sup>12</sup>. Unexpectedly, when tested with hand-held stimuli, most peripheral PC cells (34/53, 64%) displayed overt red–green opponent responses. Quantitative analysis of 35 peripheral PC cells revealed 28 (80%) in which response amplitude to isoluminant red–green exchange exceeded the response amplitude for luminance change (produced by in-phase modulation of the red and green LEDs). This is an unambiguous sign of cone opponency<sup>16</sup>. Average red–green modulation sensitivity of opponent PC cells with receptive fields above 30 degrees eccentricity ( $0.836 \pm 0.53$  (mean  $\pm$  s.d.),  $n = 11$ ; Fig. 2a, b) was not significantly different from the average value for a sample of 18 foveal cells ( $1.207 \pm 0.46$ )<sup>12</sup>. As expected<sup>1,2</sup>, however, the average sensitivity for three human observers, viewing the same



**Figure 1** Morphology of PC cells. **a**, Whole-mount view (modified with permission from ref. 9). **b**, Outline of the dendritic field of the peripheral PC cell, which encompasses 27 cones (grey). Sample box  $200 \times 200 \mu\text{m}$ . **c**, Diagram of the eye showing the recording configuration. **d**, Peri-stimulus time histograms of a PC cell at 36.5 degrees eccentricity. Responses (impulses  $\text{sec}^{-1}$ ) to red–green isoluminant exchange (top) exceed responses to luminance modulation (bottom).

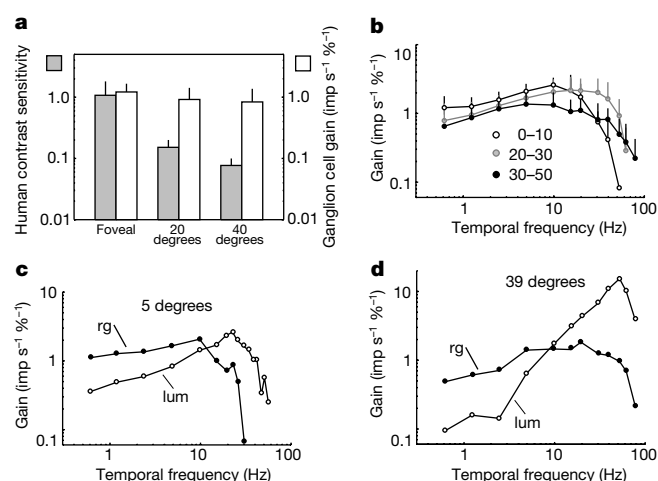
stimulus, showed a more than 10-fold decrease over this eccentricity range (from  $1.073 \pm 0.74$  at the fovea, to  $0.077 \pm 0.02$  at 30 degrees; Fig. 2a).

Temporal frequency modulation transfer functions were similar in central and peripheral opponent PC cells (Fig. 2b–d): chromatic sensitivity exceeds luminance sensitivity at temporal frequencies below 10 Hz, and is lower than luminance sensitivity for higher temporal frequencies<sup>12,19,20</sup>. Finally, cell responses to red and green LEDs presented at different relative phases and temporal frequencies<sup>16,20</sup> were consistent with selective M or L cone centres for most cells in peripheral retina, although there was slightly greater variability in cone weighting and temporal parameters of peripheral PC cells when compared with their foveal counterparts.

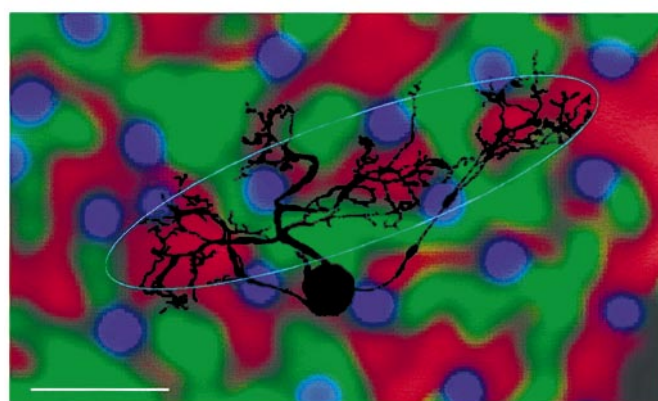
A minority (7/35, 20%) of PC cells in peripheral retina responded weakly, or not at all, to red–green isoluminant modulation. Their centre radii ( $0.228 \pm 0.128$  degrees,  $n = 5$ ) were slightly larger than those of opponent PC cells ( $0.165 \pm 0.095$  degrees,  $n = 16$ ), but this difference was not significant ( $P = 0.24$ ,  $t$ -test). No red–green

opponent cells with large, spatially non-opponent ('type 2')<sup>10</sup> receptive fields were encountered. As expected<sup>10,17,18,21</sup>, both opponent and non-opponent PC cells had significantly smaller ( $P < 0.02$ ,  $t$ -test) receptive field centre radii than phasic, magnocellular pathway cells ( $0.371 \pm 0.145$  degrees,  $n = 31$ ). These values are consistent with midget and parasol cell dendritic field radii at the eccentricity range measured<sup>8,9,22</sup>.

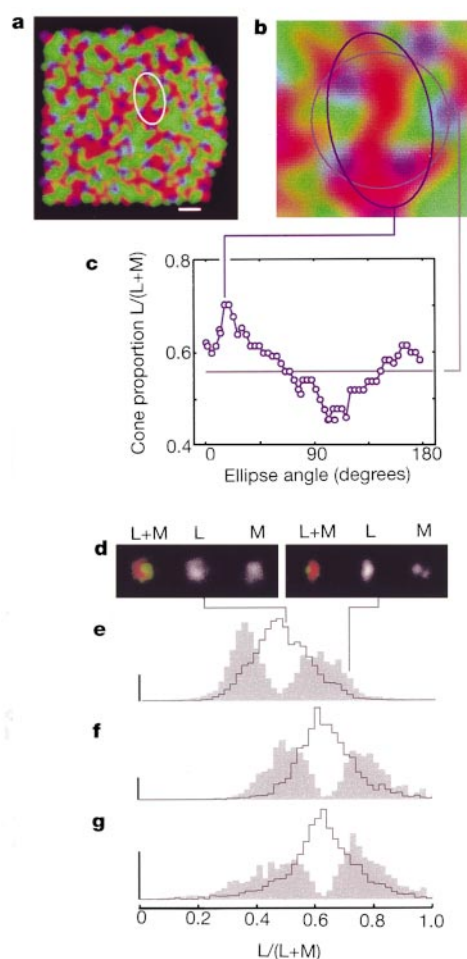
Non-opponent response characteristics have been reported for midget cells recorded *in vitro* in far peripheral retina, above 50 degrees eccentricity<sup>22</sup>, where convergence of cone input is already present at the level of midget bipolar cells. By contrast, at eccentricities below 50 degrees there is little convergence of cone inputs to midget bipolar cells<sup>13</sup>, so cone-specific signals are available to PC cell dendrites at these eccentricities. Is the morphology of these PC cells compatible with cone selectivity? A well-known feature of PC cell receptive and dendritic fields in Old World primates is that they are not circular<sup>8,23,24</sup>. Measurement of dendritic field dimensions for a sample of peripheral PC cells gave an average ellipse aspect ratio of



**Figure 2** Modulation sensitivity. **a**, Comparison of psychophysical sensitivity of three human observers with opponent PC cell contrast gain to isoluminant red–green modulation. **b**, Temporal modulation transfer functions (tMTF) of opponent PC cells for red–green isoluminant modulation. At least nine cells in each eccentricity range are shown. **c**, **d**, Chromatic (rg) and luminance (lum) tMTFs for sample foveal (**c**) and peripheral (**d**) PC cells show comparable changes in relative sensitivity to the two stimuli with temporal frequency, but the response extends to higher temporal frequencies in the peripheral retina. Error bars in **a** and **b** indicate standard deviations.



**Figure 3** PC cell dendritic tree<sup>9</sup> on a fragment of macaque cone mosaic<sup>25</sup> scaled to local cone density and gaussian filtered. Superposition of ganglion cell dendrites and the L cone mosaic was found by inspection. Local clusters in the dendritic field can be matched to the size of clusters of like (M or L) cones in the mosaic. Equal area ellipse fit to the dendritic field is superimposed. Scale bar, 50  $\mu$ m.



**Figure 4** Simulation of cone inputs to peripheral PC cells. **a**, Macaque cone photoreceptor matrix<sup>25</sup>, scaled to local cone density (710 cones per  $\text{mm}^2$ )<sup>9</sup> and Gaussian filtered. Scale bar, 0.25 degrees. **b**, Comparison of circular and elliptical sample apertures superimposed on the cone mosaic. **c**, Variation in cone weight with ellipse orientation at this position (points) compared with that for a circular aperture (solid line). Aperture radius 0.13 degrees; ellipse aspect ratio 1.8. **d**, simulated receptive field centres showing combined (coloured) and individual (greyscale) input from M and L cones. Left, circular aperture; right, optimally oriented ellipse from **b**. **e**–**g**, Distribution of cone weights. Open histograms, distribution for circular apertures; shaded histograms, distribution of optimally oriented ellipses. **e**, Randomized macaque matrix<sup>25</sup>; **f**, macaque matrix<sup>25</sup>; **g**, human subject AN matrix<sup>15</sup>. Vertical scale bar, 200 observations.

1.87. Furthermore, superposition of PC cell dendritic fields with the peripheral cone mosaic (Fig. 3) shows the potential for a high degree of selectivity<sup>8</sup>. We estimated the extent to which the anisotropy in PC cell dendritic fields might influence the balance of cone input, by simulating the connectivity of the peripheral primate retina.

The spatial distribution of cones was taken from the matrices described for the living human and macaque retina<sup>15,25</sup>. We also generated and tested matrices with random M and L cones (Fig. 4a). We compared M and L cone weight obtained from a randomly placed circular sampling aperture with that obtained from an elliptical aperture of variable orientation (Fig. 4b). This showed that even a small circular anisotropy could be the basis of a large change in the ratio of M and L cone input to individual PC cells (Fig. 4c, d).

The ellipse orientation for each simulated cell was optimized to give the greatest segregation of M or L cones. The distribution of cone weight in a large sample of such centres shows a clear bipartite distribution, with a minimum near the average M:L cone ratio, for all the cone mosaics tested (Fig. 4e–g). This is consistent with our physiological finding that most PC cell centres are dominated by either M or L cone input. By contrast, the distribution of circular apertures (or randomly oriented ellipses) shows the expected gaussian form, with a maximum near the average M:L cone ratio for each cone matrix tested (Fig. 4e–g).

The simulations show that dendritic field anisotropy in PC cells might be a sign of cone specificity. The fact that the dendrites of PC cells in human and macaque retina occupy non-overlapping domains<sup>8</sup>, which are not preferentially oriented towards the fovea (refs 22, 24; and P.R.M., K. Ghosh, A. Goodchild, unpublished data) is also consistent with this interpretation. The mechanism by which the required match of the dendritic field to clusters of like-type cones would arise is unknown. One possibility is that correlated signals from bipolar cells are a selective stimulus for synapse formation and dendritic growth by PC cells during development. This would lead the dendrites of PC cells to conform to those midget bipolar cells which, in turn, make contact with cones of the same (M or L) type. This idea is supported by evidence that activation of NMDA (N-methyl-D-aspartate) receptors leads to dendritic remodelling in ganglion cells<sup>26</sup>.

The expected loss of opponent signals in PC cells outside the fovea has been used to argue<sup>6,7</sup> that a non-midget, red–green opponent cell type must underlie the residual colour vision capacity<sup>1,5</sup> of peripheral vision. Our results strike down this objection to considering PC cells as the subcortical origin of the red–green channel for colour vision. □

## Methods

### Physiology

Adult macaque monkeys ( $n = 5$ ) were anaesthetized and intra-ocular recordings were made using standard techniques<sup>12</sup>. We recorded from a total of 131 cells at eccentricities above 20 degrees: 58 (44%) of these were phasic, magnocellular pathway cells; 54 (41%) were tonic, PC cells; 19 (14%) were tonic, blue-ON cells. We measured antidromic activation latency from the optic chiasm, to ensure that achromatic PC cells could not be mistaken for magnocellular cells<sup>11</sup>. Antidromic latencies fell into three classes: all cells with magnocellular-cell-like properties had latencies close to 5 ms; a second group (PC cells) had latencies of 8–11 ms; and blue-ON cells had latencies around 7 ms. Chromatic sensitivity was measured using luminance-matched LEDs with dominant wavelengths of 639 and 554 nm and time-averaged illuminance close to 2,000 photopic Trolands in maxwellian view<sup>12</sup>. The stimulus subtended 4.7 degrees. Response amplitude is taken as the first Fourier harmonic amplitude at the stimulus frequency. Contrast is expressed as LED modulation relative to maximum. In terms of cone contrast, 100% isoluminant red–green modulation produced 68% Michelson contrast in the M cone and 20% Michelson contrast in the L cone<sup>16</sup>. Receptive field dimensions were measured from responses to drifting sinusoidal luminance gratings<sup>17</sup>.

### Psychophysics

Three observers (P.R.M., S.G.S. and subject V.B., who did not know the purpose of the experiment) set contrast detection thresholds by the method of adjustment. Stimulus parameters were identical to those for physiological measurements except that the temporal modulation frequency was 1.024 Hz for psychophysics and 1.22 Hz for physiology.

### Anatomy

Ganglion cells were filled with Neurobiotin in an *in vitro* preparation of macaque retina<sup>9,22</sup>. The area equivalent ellipse was calculated from a polygon drawn to connect the outermost points of the dendritic field for 98 PC and magnocellular cells. The ellipse aspect ratio for PC cells ( $1.87 \pm 0.64$ ) was significantly greater ( $P < 0.01$ , *t*-test) than that for magnocellular cells ( $1.38 \pm 0.26$ ).

### Modelling

Cone centre position coordinates for macaque M5 (ref. 25) and subject AN (ref. 15) were assigned to a bit-plane image matrix and masked with circular or elliptical apertures. For the L/M random cone matrix the same centre coordinates were used but the identity of each L or M cone was randomly reassigned with equal probability. Aperture dimensions (0.1–0.2 degree radius, 1.5–1.8 aspect ratio) are compatible with measurements of PC-receptive<sup>17,23</sup> and dendritic fields<sup>8,9,22,24</sup> (Fig. 3), and encompass 15–25 cones at 35 degrees. The L and M cones in each aperture were counted using feature-based image detection algorithms (MathWorks, Natick MA, USA). Optimal ellipse orientation was determined using a simple 'hebbian' rule: the dominant (M or L cone) was measured in an initial, circular estimate of the receptive field composition, and then the orientation was chosen that maximized this initial bias. The orientation step size was 22.5 degrees. For illustration purposes, we filtered the cone mosaic (Figs 3 and 4a, b) with a two-dimensional gaussian (radius 0.030 degrees)<sup>27</sup>. Aperture dimensions (radius in degrees and aspect ratio) for histograms were 0.13 and 1.8 (Fig. 4e); and 0.16 and 1.6 (Fig. 4f); 0.16 and 1.6 (Fig. 4g). Cone weight histograms were filtered with 2% uniform noise to reduce artifact from quantization.

Received 20 October 2000; accepted 9 January 2001.

- Noorlander, C., Koenderink, J. J., Den Ouden, R. J. & Edens, B. W. Sensitivity to spatiotemporal colour contrast in the peripheral visual field. *Vision Res.* **23**, 1–11 (1983).
- Mullen, K. T. Colour vision as a post-receptoral specialization of the central visual field. *Vision Res.* **31**, 119–130 (1991).
- Shapley, R. & Perry, V. H. Cat and monkey retinal ganglion cells and their visual functional roles. *Trends Neurosci.* **9**, 229–235 (1986).
- Lennie, P., Haake, P. W. & Williams, D. R. in *Computational Models of Visual Processing* (eds Landy, M. S. & Movshon, J. A.) 71–82 (MIT Press, Cambridge, MA, 1991).
- Mullen, K. T. & Kingdom, F. A. A. Losses in peripheral colour sensitivity predicted from "hit and miss" post-receptoral cone connections. *Vision Res.* **36**, 1995–2000 (1996).
- Calkins, D. J. & Sterling, P. Evidence that circuits for spatial and color vision segregate at the first retinal synapse. *Neuron* **24**, 313–321 (1999).
- Rodieke, R. W. in *From Pigments to Perception: Advances in Understanding Visual Processes* (eds Valberg, A. & Lee, B. B.) 83–93 (Plenum, London, 1991).
- Dacey, D. M. The mosaic of midget ganglion cells in the human retina. *J. Neurosci.* **13**, 5334–5355 (1993).
- Goodchild, A. K., Ghosh, K. K. & Martin, P. R. Comparison of photoreceptor spatial density and ganglion cell morphology in the retina of human, macaque monkey, cat, and the marmoset *Callithrix jacchus*. *J. Comp. Neurol.* **366**, 55–75 (1996).
- Wiesel, T. N. & Hubel, D. H. Spatial and chromatic interactions in the lateral geniculate body of the rhesus monkey. *J. Neurophysiol.* **29**, 1115–1156 (1966).
- De Monasterio, F. M. & Gouras, P. Functional properties of ganglion cells of the rhesus monkey retina. *J. Physiol. (Lond.)* **251**, 167–195 (1975).
- Lee, B. B., Pokorny, J., Smith, V. C., Martin, P. R. & Valberg, A. Luminance and chromatic modulation sensitivity of macaque ganglion cells and human observers. *J. Opt. Soc. Am.* **7**, 2223–2236 (1990).
- Wässle, H., Grünert, U., Martin, P. R. & Boycott, B. B. Immunocytochemical characterization and spatial distribution of midget bipolar cells in the macaque monkey retina. *Vision Res.* **34**, 561–579 (1994).
- Packer, O. S., Williams, D. R. & Bensinger, D. G. Photopigment transmittance imaging of the primate photoreceptor mosaic. *J. Neurosci.* **16**, 2251–2260 (1996).
- Roorda, A. & Williams, D. R. The arrangement of the three cone classes in the living human eye. *Nature* **397**, 520–522 (1999).
- Smith, V. C., Lee, B. B., Pokorny, J., Martin, P. R. & Valberg, A. Responses of macaque ganglion cells to the relative phase of heterochromatically modulated lights. *J. Physiol. (Lond.)* **458**, 191–221 (1992).
- Derrington, A. M. & Lennie, P. Spatial and temporal contrast sensitivities of neurones in lateral geniculate nucleus of macaque. *J. Physiol. (Lond.)* **357**, 219–240 (1984).
- Dreher, B., Fukada, Y. & Rodieke, R. W. Identification, classification and anatomical segregation of cells with X-like and Y-like properties in the lateral geniculate nucleus of Old-World primates. *J. Physiol. (Lond.)* **258**, 433–452 (1976).
- Gouras, P. & Zrenner, E. Enhancement of luminance flicker by color-opponent mechanisms. *Science* **205**, 587–589 (1979).
- Lankheet, M. J. M., Lennie, P. & Krauskopf, J. Temporal–chromatic interactions in LGN P-cells. *Visual Neurosci.* **15**, 47–54 (1998).
- Kaplan, E., Lee, B. B. & Shapley, R. M. in *New Views of Primate Retinal Function* Vol. 9 (eds Osborne, N. & Chader, J.) 273–336 (Pergamon, New York, 1989).
- Dacey, D. M. Primate retina: cell types, circuits and color opponency. *Prog. Retin. Res.* **18**, 737–763 (1999).
- Smith, E. L., Chino, Y. M., Ridder, W. H., Kitagawa, K. & Langston A. Orientation bias of neurons in the lateral geniculate nucleus of macaque monkeys. *Visual Neurosci.* **5**, 525–545 (1990).
- Watanabe, M. & Rodieke, R. W. Parasol and midget ganglion cells of the primate retina. *J. Comp. Neurol.* **289**, 434–454 (1989).
- Roorda, A., Metha, A., Lennie, P. & Williams, D. R. Packing arrangement of the three cone classes in primate retina. *Vision Res.* (in the press).
- Wong, W. T., Faulkner-Jones, B. E., Sanes, J. R. & Wong, R. O. Rapid dendritic remodeling in the developing retina: dependence on neurotransmission and reciprocal regulation by Rac and Rho. *J. Neurosci.* **20**, 5024–5036 (2000).
- Navarro, R., Artal, P. & Williams, D. R. Modulation transfer of the human eye as a function of retinal eccentricity. *J. Opt. Soc. Am.* **10**, 201–212 (1993).



# Acknowledgements

We thank B. Roser and B. Grünert for technical assistance; B. Dreher, U. Grünert, P. Lennie and H. Wässle for helpful discussions and suggestions; A. Goodchild and K. Ghosh for Neurobiotin labelling of ganglion cells; and A. Roorda for cone matrix data.

Correspondence and requests for materials should be addressed to P.M. (e-mail: paulm@physiol.usyd.edu.au).

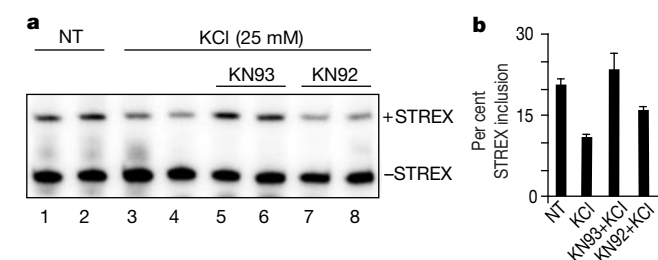
## A CaMK IV responsive RNA element mediates depolarization-induced alternative splicing of ion channels

Jiuyong Xie\* & Douglas L. Black\*†

\* Howard Hughes Medical Institute, and † Department of Microbiology & Molecular Genetics, University of California Los Angeles, California 90095-1662, USA

Calcium regulation of gene expression is critical for the long-lasting activity-dependent changes in cellular electrical properties that underlie important physiological functions such as learning and memory<sup>1</sup>. Cellular electrical properties are diversified through the extensive alternative splicing of ion channel pre-messenger RNAs<sup>2</sup>; however, the regulation of splicing by cell signalling pathways has not been well explored. Here we show that depolarization of GH<sub>3</sub> pituitary cells represses splicing of the STREX exon<sup>3</sup> in BK potassium channel transcripts through the action of Ca<sup>2+</sup>/calmodulin-dependent protein kinases (CaMKs). Overexpressing constitutively active CaMK IV, but not CaMK I or II, specifically decreases STREX inclusion in the mRNA. This decrease is prevented by mutations in particular RNA repressor sequences. Transferring 54 nucleotides from the 3' splice site upstream of STREX to a heterologous gene is sufficient to confer CaMK IV repression on an otherwise constitutive exon. These experiments define a CaMK IV-responsive RNA element (CaRRE), which mediates the alternative splicing of ion channel pre-mRNAs. The CaRRE presents a unique molecular target for inducing long-term adaptive changes in cellular electrical properties. It also provides a model system for dissecting the effect of signal transduction pathways on alternative splicing.

BK (Slo) channels participate in the repolarization and fast after-hyperpolarization of action potentials and are important in setting the firing properties of neurons<sup>4,5</sup>. Their diverse properties result largely from extensive alternative splicing of the *Slo* transcript<sup>6–11</sup>.



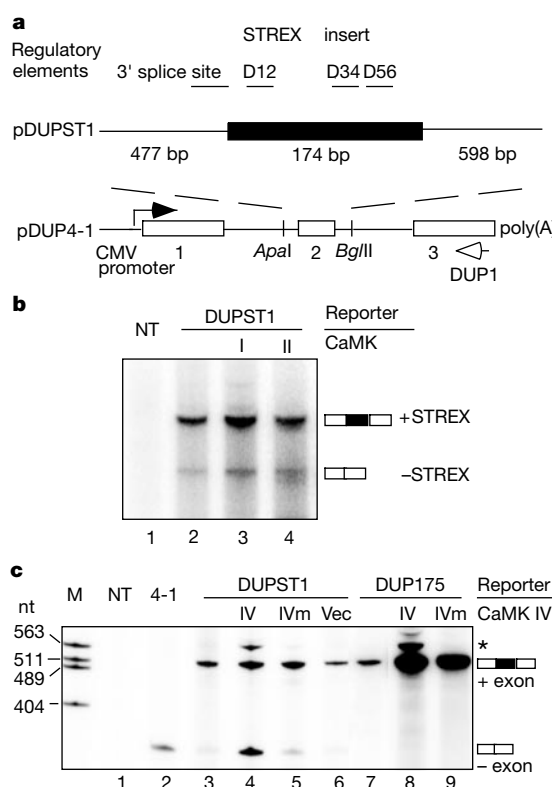
**Figure 1** Depolarization-induced changes in STREX splicing depend on CaMK. **a**, Denaturing gel electrophoresis of Slo RT-PCR products from GH<sub>3</sub> cells, including (+) or excluding (–) the STREX exon, 6 h after treatment with nothing (NT, lanes 1 and 2), KCl (lanes 3 and 4), KCl plus KN93 (lanes 5 and 6), or KCl plus KN92 (lanes 7 and 8). Two independent cell samples for each treatment are shown. **b**, Percentage of mRNA containing STREX after the treatments shown in **a**. Results are mean  $\pm$  s.d.,  $n = 3$ .

The STREX exon of Slo confers higher Ca<sup>2+</sup> sensitivity on the channel<sup>3,11–14</sup>, and may contribute to the repetitive firing of adrenal chromaffin cells<sup>3,5</sup> and the tuning of hearing frequencies in cochlear hair cells<sup>11,13</sup>. The differential expression of STREX suggests that its splicing is modulated by many physiological factors<sup>3,11,15</sup>.

Polymerase chain reaction after reverse transcription (RT-PCR) of Slo mRNA indicated that GH<sub>3</sub> rat pituitary cells include the STREX exon in about 20% of their endogenous Slo transcripts (Fig. 1a, lanes 1, 2, and Fig. 1b). When these cells were depolarized by adding KCl (25 mM), STREX splicing was reduced by 50% in 6 h (lanes 3, 4). Adding the CaMK inhibitor KN93 to the media, before the KCl depolarization, completely blocked this decrease (30  $\mu$ M, lanes 5, 6). In contrast, the inactive analogue KN92 had only a slight effect (lanes 7, 8). Thus, depolarization of GH<sub>3</sub> cells represses STREX exon splicing, and this repression apparently occurs through a CaMK-dependent pathway.

CaMKs I, II and IV can be activated by cell depolarization and inhibited by KN93. CaMK IV, in particular, is important in depolarization-regulated gene expression<sup>16</sup>. Because all three enzymes are expressed in GH<sub>3</sub> cells (Supplementary Information Fig. 1a), any one of them might be involved in the repression of STREX splicing.

We wanted to determine whether the regulation was acting through the sequences surrounding the STREX exon itself and also was not an effect of transcriptional regulation or RNA degradation. The mouse STREX exon with partial flanking intron sequences



**Figure 2** CaMK IV overexpression represses STREX exon inclusion in HEK cells. **a**, The STREX splicing construct pDUPST1. The STREX insert replaces DUP4-1 sequences between the *Apal* and *Bgl*II sites. The CMV promoter, primer extension primer DUP1 (arrows), exons (boxes), introns (lines) and regulatory elements are indicated. **b**, Primer extension assays of pDUPST1 co-transfected with CaMK I-dCT (I) or CaMK II-dCT (II). **c**, Primer extension assay as in **b** with CaMK IV. M, nucleotide size marker; NT, mock transfection (lane 1); 4-1, pDUP4-1 (lane 2); IV, CaMK IV-dCT (lanes 4 and 8); IVm, CaMK IV-dCT75E (lanes 5 and 9); Vec, pcDNA3.1(+) (lane 6). Asterisk indicates apparently early terminated products from unspliced mRNA.

was cloned between constitutive  $\beta$ -globin exons of plasmid DUP4-1 (ref. 17; Fig. 2a). This clone (DUPST1) was transfected into HEK293 cells, and the expressed mRNAs were assayed for STREX exon inclusion by primer extension.

The STREX exon was strongly included in the DUPST1 mRNA in transfected HEK cells ( $\sim 90\%$ ; Fig. 2c, lane 3). To test whether the exon in the transfected gene was regulated by CaMKs, we co-transfected DUPST1 with plasmids expressing constitutively active forms of CaMK I, II or IV (CaMK I-dCT (ref. 18), II-dCT (ref. 19) or IV-dCT (ref. 20)). Co-expressed CaMK I or II did not affect STREX splicing (Fig. 2b); however, CaMK IV-dCT strongly reduced STREX splicing ( $\sim 60\%$  inclusion; Fig. 2c, lane 4). In contrast, neither a kinase-dead mutant, CaMK IV-dCTK75E (ref. 21; lane 5), nor the empty vector pcDNA3.1(+) (lane 6) affected STREX splicing. The activity of the expressed kinases was confirmed by measuring phosphorylation of CRE-binding protein

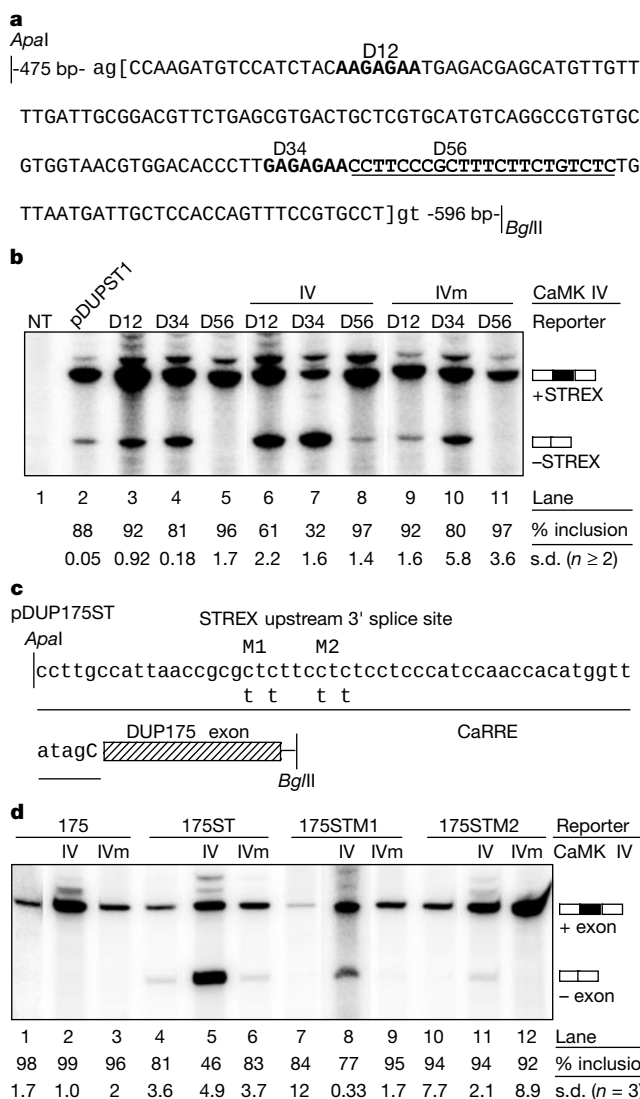
at Ser 133 (Supplementary Information Fig. 1b). The splicing repression thus seems to be specific for CaMK IV-dCT.

The DUPST1 transcript level varied between some lanes of the gels in Fig. 2 (see Fig. 3b). We controlled for the effect of expression level on splicing by measuring STREX inclusion relative to the overall gene transcription compared with a co-transfected control plasmid. STREX inclusion was constant at all observed levels of expression (data not shown). In addition, CaMK IV-dCT did not reduce the splicing of a 175-base-pair (bp) constitutive exon of pDUP175 derived from  $\beta$ -globin<sup>22</sup> (Fig. 2c, lanes 7–9). Because the DUPST1 and DUP175 constructs have the same CMV promoter, flanking exons and poly(A) signal, it is unlikely that the change in mRNA products is due to differences in transcriptional regulation. The regulation by CaMK IV is a property of the exon itself, and not of the transcription unit of which it is a part.

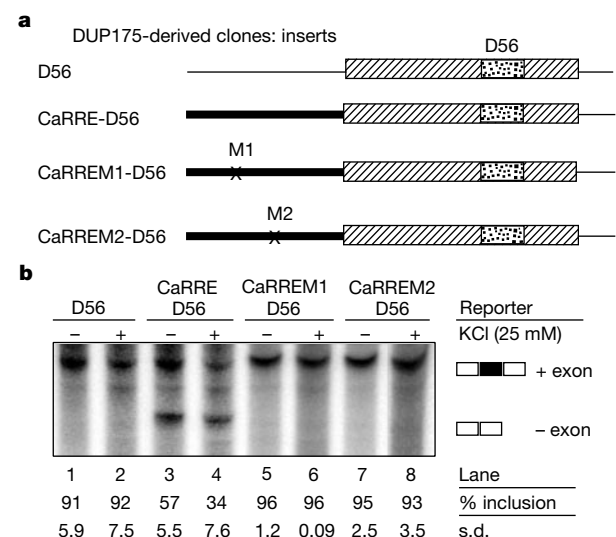
To identify specific features of the STREX exon that are needed for CaMK IV repression, we first made deletions in STREX exonic sequences that were similar to either purine-rich enhancer (D12 and D34) or pyrimidine-rich repressor (D56) elements identified in other alternative exons<sup>23,24</sup> (Fig. 3a). Deletion of the D12 sequence had almost no effect (Fig. 3b, lane 3), and deletion of the D34 element caused a slight decrease in exon inclusion (lane 4). The D56 deletion increased inclusion (lane 5). Most striking was the effect of these mutations on the response to CaMK IV.

CaMK IV-dCT reduced D12 mutant and wild-type exon splicing similarly (Fig. 3b, lanes 3, 6, 9; and Fig. 2c, lane 4), but induced greater repression of splicing with the D34 mutant than with the wild-type exon (lane 7). D34 exhibited 81% exon inclusion when transfected alone (lane 4), and 32% inclusion when co-transfected with CaMK IV-dCT (lane 7). Most significantly, the D56 mutant exon showed almost no repression by CaMK IV-dCT (96% inclusion transfected alone (lane 5), and 97% with CaMK IV-dCT (lane 8)). In other experiments, mutation of thymidines in the D56 element to adenosines, while maintaining the exon length, also prevented the splicing repression by CaMK IV. Thus, the D56 pyrimidine-rich element is required for CaMK IV-mediated repression of STREX splicing in this construct.

Most regulated exons that have been characterized are controlled



**Figure 3** STREX RNA elements responsive to CaMK IV repression in HEK cells. **a**, Potential splicing regulatory elements (D12, D34 and D56) in the STREX exon (in brackets), with the D56 element underlined. **b**, Primer extension assay of wild-type or mutant pDUPST1 transfected alone (lanes 2–5), or with CaMK IV-dCT (IV, lanes 6–8) or -dCTK75E (IVm, lanes 9–11). **c**, DUP175ST construct showing the STREX 3' splice site upstream of the DUP175 exon (box). The CaRRE (53-nucleotide intron + 1-nucleotide exon) and its M1 or M2 mutants are underlined. **d**, Primer extension assay of plasmids indicated at the top transfected alone (lanes 1,4,7,10), or with IV (lanes 2,5,8,11) or IVm (lanes 3,6,9,12).



**Figure 4** The CaRRE responds to depolarization in GH<sub>3</sub> cells. **a**, Diagrams of the constructs used in the transfection assays. The STREX 3' splice site (thick line), the DUP175 exon (box), the STREX D56 sequence (dotted box), and the DUP175 introns (thin lines) are indicated. **b**, Primer extension assay of RNA from GH<sub>3</sub> cells transfected with the plasmids shown in **a**, with (+) or without (–) KCl treatment. Plasmids are indicated above, and the per cent exon inclusion ( $n = 2$ ) below.

by multiple enhancer and repressor elements, which are found in flanking introns as well as the exon<sup>23,25,26</sup>. Exons controlled by pyrimidine-rich repressor elements often contain additional repressor elements in their upstream 3' splice site<sup>25,26</sup>. Further mutagenesis of the DUPST1 plasmid indicated that several intronic sequences, including the 3' splice site, were also affecting STREX splicing (Supplementary Information Fig. 2). However, the effect of point mutations in these elements was not as marked as that of the D56 mutation in pDUPST1, presumably because of redundancy in the regulatory sequences<sup>23</sup>. To avoid this redundancy, we transferred small segments from STREX to the CaMK IV non-responsive DUP175 exon, where the effect of each element could be measured in the absence of other STREX sequences. Although the D56 element is required for CaMK IV repression of the STREX exon in pDUPST1, it was not sufficient to repress splicing when transferred to the DUP175 exon (see Fig. 4b, lanes 1, 2; and data not shown). Other STREX elements were apparently needed for D56 to function.

In contrast to D56, the STREX 3' splice site had a large effect on DUP175 splicing (Fig. 3). The constitutive DUP175 exon was not responsive to CaMK IV (Fig. 3d, lanes 1–3). When the 3' splice site upstream of DUP175 was replaced with the corresponding sequence from STREX, this DUP175ST exon was strongly repressed by CaMK IV (lanes 4–6). Mutations M1 and M2 in the STREX splice site

almost eliminated CaMK IV repression, yielding a nearly constitutive exon (Fig. 3c, d, lanes 7–9 and 10–12, respectively). Another constitutive CaMK IV, activated by point mutations rather than by truncation, also repressed DUP175ST exon inclusion (Supplementary Information Fig. 3). Thus, the upstream 3' splice site of STREX is sufficient to confer CaMK IV-mediated repression on a heterologous exon, and the precise sequence in its polypyrimidine tract is critical to its function. We call this sequence a CaMK IV-responsive RNA element (CaRRE).

For the STREX exon, both D56 and the CaRRE were needed for CaMK IV repression of splicing. In contrast, with DUP175 only the CaRRE was needed. When most of the Slo intron sequences were deleted from DUPST1, splicing of STREX still responded to the CaRRE and D56 but was lower overall (see Fig. 5c, DUPST2; and Supplementary Information Fig. 2b, lanes 1, 3). This indicates that additional elements contribute to STREX regulation. Similar differences in the requirements of native and reporter exons for regulatory elements are seen in other systems<sup>23,25</sup>. It will require extensive mutagenesis to identify all the elements affecting STREX splicing; however, from our results it is clear that particular RNA elements confer CaMK responsiveness on an exon.

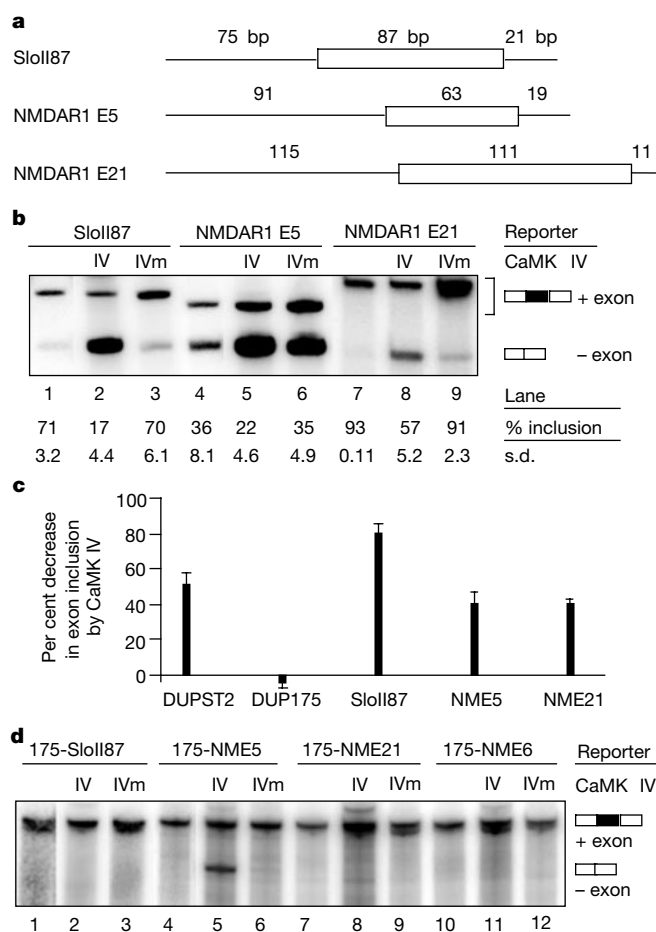
We next determined whether the CaRRE would respond to depolarization in GH<sub>3</sub> cells as it did to CaMK IV. Tests with several constructs indicated that GH<sub>3</sub> cells show even stronger STREX or DUP exon inclusion than HEK cells, making the exons less sensitive to repression by the 3' splice site. To make a more repressible CaRRE-dependent exon, we modified the DUP175 exon to contain both the 3' splice site and the D56 repressor element from STREX (Fig. 4). The DUP175-D56 exon, containing just the D56 element, was completely included in GH<sub>3</sub> cells and was not repressed by depolarization (Fig. 4b, lanes 1, 2). When the CaRRE was added to this exon, exon inclusion was reduced to about 60% in normal media (CaRRE-D56; lane 3). Most significantly, depolarization now further reduced splicing to about 30% (lane 4). Again, the M1 and M2 mutations abolished this repression (CaRRE-M1-D56 and CaRRE-M2-D56; lanes 5–8). Thus, the CaRRE confers repression by depolarization, supporting the idea that depolarization and overexpressed CaMK IV repress splicing through the same pathway.

Many other proteins involved in cellular excitation show complex splicing variation. To examine whether CaMK IV affects other exons, we cloned several variant exons into the DUP4-1 reporter and tested their splicing in the presence or absence of CaMK IV (Fig. 5). We found that another exon of Slo (SloII87; see Methods), as well as exons 5 and 21 of NMDAR1 (*N*-methyl-D-aspartate receptor 1; ref. 27) were also repressed by CaMK IV (Fig. 5b). NMDAR1 exons 5 and 21 show a splicing decrease similar to the STREX exon in DUPST2 (lanes 5, 8). SloII87 was more strongly repressed (lane 2).

To examine whether the 3' splice site is a common target of CaMK IV repression, we transferred only the 3' splice sites of these exons to DUP175. We found that the splice site of NMDAR1 exon 5 also conferred CaMK IV responsiveness (Fig. 5d, lane 5), whereas the sites from SloII87, NMDAR1 exon 21 and the constitutive NMDAR1 exon 6 had no effect (lanes 2, 8, 11). Thus, CaMK IV apparently instigates a coordinate change in the splicing of a group of variant exons. It will be interesting to test additional transcripts that affect cell excitation, and the physiological effect of these splicing changes by CaMK IV.

Pyrimidine-rich elements are often binding sites for the polypyrimidine-tract binding protein (PTB)<sup>27,28</sup>, which is implicated in the repression of a number of exons. Indeed, PTB is known to bind the 3' splice site of NMDAR1 exon 5 (ref. 27). PTB is hence an appealing target for CaMK IV; however, we have not observed any changes in PTB on GH<sub>3</sub> cell depolarization, nor have we seen the direct phosphorylation of PTB by purified CaMK IV. Thus, the identity of the CaMK IV targets requires further investigation.

In summary, we have shown that both cell depolarization and



**Figure 5** CaMK IV repression and a CaRRE in other ion channel exons. **a**, Exons SloII87, and 5 (E5) and 21 (E21) of NMDAR1 inserted into pDUP4-1. **b**, Primer extension assay of the plasmids shown in **a** transfected alone (lanes 1, 4, 7), or with CaMK IV-dCT (IV, lanes 2, 5, 8) or -dCT75E (IVm, lanes 3, 6, 9). **c**, Per cent reduction in exon inclusion by CaMK IV-dCT for DUPST2, DUP175, and the plasmids shown in **b** (mean  $\pm$  s.d.,  $n \geq 3$ ). **d**, Primer extension assay of DUP175-derived plasmids (containing 3' splice sites as in Fig. 3c) transfected alone (lanes 1, 4, 7, 10), or with IV (lanes 2, 5, 8, 11) or IVm (lanes 3, 6, 9, 12).



CaMK IV signalling regulate the alternative splicing of ion channels through a defined RNA regulatory element. Through the CaRRE, chronic input by action potentials is likely to decrease STREX inclusion, and thus decrease the  $\text{Ca}^{2+}$  sensitivity and opening probability of BK channels. This, consequently, will alter the repolarization, after-hyperpolarization, and firing properties of later waves of action potentials. Thus, the CaRRE provides a molecular target for controlling activity-driven long-lasting changes in cellular electrical properties. In cerebellar neurons, where STREX is expressed<sup>15</sup>, the late phase of long-term depression requires both transcription and CaMK IV (ref. 29). It will be interesting to determine whether the regulated splicing of Slo is also involved.

Other adaptive changes that might result from regulated BK channel splicing include neurotransmitter release and hormone secretion. In addition to the exons tested here, the large number of splice variants of ion channels, and the widespread alternative splicing in neuronal and endocrine cells in general, provide many likely targets for  $\text{Ca}^{2+}$  signalling. Thus, this regulation is an attractive molecular mechanism for the fine-tuning of cellular electrical properties<sup>30</sup>. The STREX exon also provides a working system to study the effects of cell signalling on splicing and to identify the molecules involved. □

## Methods

### Cloning genomic DNA encoding regulated exons

Genomic DNA containing the mouse STREX exon with partial flanking introns was subcloned from the mSlo bacterial artificial chromosome (BAC) clone 75D24, identified from the mouse BAC library (Genome Systems) using the *HindIII*–*XhoI* fragment of Slo cDNA as a probe<sup>3</sup>. The partial intronic sequences surrounding the 87-bp exon at alternative splicing site II of Slo (thus named SloII87, encoding a 29-amino-acid insert<sup>8</sup>) were obtained with the mouse DNA Genomic Walker kit (Clontech). We amplified NMDAR1 (*N*-methyl-D-aspartate receptor NR1) exons 5 and 21 by PCR from human genomic DNA using Pfu DNA polymerase, on the basis of GenBank genomic sequences Z32773 and Z32774.

### Plasmid construction

Exons with flanking intron sequences were cloned between the *Apal* and *BglII* sites of pDUP4-1 (ref. 17). Mutants were made by PCR using Pfu DNA polymerase. CaMK IV-dCT and CaMK IV-dCTK75E inserts with Flag tags at the amino termini<sup>20,21</sup> were cloned into pcDNA3.1(+) at the *BamHI* site. CaMK II-dCT (residues 1–290)<sup>19</sup> was cloned into pcDNA3.1(+) with a Flag tag at the N terminus. pDUP175 is the same as pDUP171 (ref. 22), except that the middle exon and the partial flanking introns were recombined into pDUP4-1 between the *Apal* and the *BglII* site<sup>17</sup>, to give it the same restriction sites as the other splicing constructs. We confirmed all constructs by sequencing. The translation stop codons for pDUPST1, pDUPST1-D34 and pDUPST1-D56 mRNA are the same; thus, their differences are unlikely to be caused by nonsense-mediated RNA decay.

### Cell culture, treatments and transfections

We cultured GH<sub>3</sub> cells in HAM F10 with 10% horse serum plus 5% fetal bovine serum and HEK cells in DMEM with 15% calf serum. For the depolarization of GH<sub>3</sub> cells, overnight cultures were treated by adding KCl to 25 mM (plus 3.8 mM in the F10 media), with or without CaMK inhibitor KN93 (30  $\mu\text{M}$ ) or the inactive analogue KN92 (30  $\mu\text{M}$ ). Transfections were done with SuperFect (Qiagen) using overnight cultures at about 60% confluency. For HEK cells, 2  $\mu\text{g}$  of splicing plasmid and/or 6  $\mu\text{g}$  of CaMK plasmid, its mutant, or a vector control were used per well of a six-well plate. For transfections in GH<sub>3</sub> cells, 12  $\mu\text{g}$  of splicing reporter plasmid was used for each 90-mm plate. KCl was added 3 h after transfection and maintained thereafter. After transfection, RNA was extracted at 16–20 h for HEK cells, or 2 days for GH<sub>3</sub> cells.

### RT-PCR and primer extension

We performed RT-PCR as described<sup>3</sup>, except that the RT product was amplified for only 25 cycles, a <sup>32</sup>P-labelled downstream primer was included, and the PCR products were quantified by phosphorimager. The ratios of the bands were constant as monitored from 22 to 26 cycles, suggesting that the amplification was in linear phase. We found that cell lines LA-N-5, PC12, AtT20 and GH<sub>3</sub> had partial inclusion of STREX. Primer extensions were done as described<sup>17</sup>, except that the primer was DUP1 (sequence: 5'-GCAGCTCACTCAGTGTGGCA-3') and the annealing temperature was 55 °C. This primer gives rise to a 342 bp primer extension product when the middle exon from pDUP4-1 is not included in HEK cells. The length of the test exon-included product should be 341 plus the test exon length.

Received 12 December 2000; accepted 5 February 2001.

1. Finkbeiner, S. & Greenberg, M. E.  $\text{Ca}^{2+}$  channel-regulated neuronal gene expression. *J. Neurobiol.* **37**, 171–189 (1998).

2. Coetzee, W. A. *et al.* Molecular diversity of  $\text{K}^{+}$  channels. *Ann. N. Y. Acad. Sci.* **868**, 233–285 (1999).
3. Xie, J. & McCobb, D. P. Control of alternative splicing of potassium channels by stress hormones. *Science* **280**, 443–446 (1998).
4. Hille, B. *Ionic Channels of Excitable Membranes* (Sinauer Associates, Sunderland, MA, 1992).
5. Lingle, C. J., Solaro, C. R., Prakriya, M. & Ding, J. P. Calcium-activated potassium channels in adrenal chromaffin cells. *Ion Channels* **4**, 261–301 (1996).
6. Butler, A., Tsunoda, S., McCobb, D. P., Wei, A. & Salkoff, L. MSlo, a complex mouse gene encoding “maxi” calcium-activated potassium channels. *Science* **261**, 221–224 (1993).
7. Lagrutta, A., Shen, K.-Z., North, R. A. & Adelman, J. P. Functional differences among alternatively spliced variants of Slowpoke, a *Drosophila* calcium-activated potassium channel. *J. Biol. Chem.* **269**, 20347–20351 (1994).
8. Tseng-Crank, J. *et al.* Cloning, expression, and distribution of functionally distinct  $\text{Ca}^{2+}$ -activated  $\text{K}^{+}$  channel isoforms from human brain. *Neuron* **13**, 1315–1330 (1994).
9. Navaratnam, D. S., Bell, T. J., Tu, T. D., Cohen, E. L. & Oberholzer, J. C. Differential distribution of  $\text{Ca}^{2+}$ -activated  $\text{K}^{+}$  channel splice variants among hair cells along the tonotopic axis of the chick cochlea. *Neuron* **19**, 1077–1085 (1997).
10. Rosenblatt, K. P., Sun, Z.-P., Heller, S. & Hudspeth, A. J. Distribution of  $\text{Ca}^{2+}$ -activated  $\text{K}^{+}$  channel isoforms along the tonotopic gradient of the chicken's cochlea. *Neuron* **19**, 1061–1075 (1997).
11. Jones, E. M. C., Gray-Keller, M. & Fettiplace, R. The role of  $\text{Ca}^{2+}$ -activated  $\text{K}^{+}$  channel spliced variants in the tonotopic organization of the turtle cochlea. *J. Physiol. (Cambridge)* **518**, 653–665 (1999).
12. Saito, M., Nelson, C., Salkoff, L. & Lingle, C. J. A cysteine-rich domain defined by a novel exon in a Slo variant in rat adrenal chromaffin cells and PC12 cells. *J. Biol. Chem.* **272**, 11710–11717 (1997).
13. Ramanathan, K., Michael, T. H., Jiang, G.-J., Hiel, K. & Fuchs, P. A. A molecular mechanism for electrical tuning of cochlear hair cells. *Science* **283**, 215–217 (1999).
14. Shipston, M. J., Duncan, R. R., Clark, A. G., Antoni, F. A. & Tian, L. Molecular components of large conductance calcium-activated potassium (BK) channels in mouse pituitary corticotropes. *Mol. Endocrinol.* **13**, 1728–1737 (1999).
15. Ferrer, J., Wasson, J., Salkoff, L. & Permutt, M. A. Cloning of human pancreatic islet large conductance  $\text{Ca}^{2+}$ -activated  $\text{K}^{+}$  channel (hSlo) cDNAs: Evidence for high levels of expression in pancreatic islets and identification of a flanking genetic marker. *Diabetologia* **39**, 891–898 (1996).
16. Soderling, T. R. The  $\text{Ca}^{2+}$ -calmodulin-dependent protein kinase cascade. *Trends Biochem. Sci.* **24**, 232–236 (1999).
17. Modafferi, E. F. & Black, D. L. A complex intronic splicing enhancer from the *c-src* pre-mRNA activates inclusion of a heterologous exon. *Mol. Cell. Biol.* **17**, 6537–6545 (1997).
18. Sun, P., Lou, L. & Maurer, R. A. Regulation of activating transcription factor-1 and the cAMP response element-binding protein by  $\text{Ca}^{2+}$ /calmodulin-dependent protein kinases type I, II, and IV. *J. Biol. Chem.* **271**, 3066–3073 (1996).
19. Sun, P., Enslin, H., Myung, P. S. & Maurer, R. A. Differential activation of CREB by  $\text{Ca}^{2+}$ /calmodulin-dependent protein kinases type II and type IV involves phosphorylation of a site that negatively regulates activity. *Genes Dev.* **8**, 2527–2539 (1994).
20. Chatila, T., Anderson, K. A., Ho, N. & Means, A. R. A unique phosphorylation-dependent mechanism for the activation of  $\text{Ca}^{2+}$ /calmodulin-dependent protein kinase type IV/GR. *J. Biol. Chem.* **271**, 21542–21548 (1996).
21. Miranti, C. K., Ginty, D. D., Huang, G., Chatila, T. & Greenberg, M. E. Calcium activates serum response factor-dependent transcription by a Ras- and Elk-1-independent mechanism that involves a  $\text{Ca}^{2+}$ /calmodulin-dependent kinase. *Mol. Cell. Biol.* **15**, 3672–3684 (1995).
22. Dominski, Z. & Kole, R. Selection of splice sites in pre-mRNAs with short internal exons. *Mol. Cell. Biol.* **11**, 6075–6083 (1991).
23. Modafferi, E. F. & Black, D. L. Combinatorial control of a neuron-specific exon. *RNA* **5**, 687–706 (1999).
24. Lopez, A. J. Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. *Annu. Rev. Genet.* **32**, 279–305 (1998).
25. Smith, C. & Valcarcel, J. Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem. Sci.* **25**, 381–388 (2000).
26. Zhang, L., Ashiya, M., Sherman, T. G. & Grabowski, P. J. Essential nucleotides direct neuron-specific splicing of gamma-2 pre-mRNA. *RNA* **2**, 682–698 (1996).
27. Zhang, L., Liu, W. & Grabowski, P. J. Coordinate repression of a trio of neuron-specific splicing events by the splicing regulator PTB. *RNA* **5**, 117–130 (1999).
28. Chou, M.-Y., Underwood, J. G., Nikolic, J. M., Luu, M. H. T. & Black, D. L. Multisite RNA binding and release of polypyrimidine tract binding protein during the regulation of *c-src* neural-specific splicing. *Mol. Cell* **5**, 949–957 (2000).
29. Ahn, S., Ginty, D. D. & Linden, D. J. A late phase of cerebellar long-term depression requires activation of CaMKIV and CREB. *Neuron* **23**, 559–568 (1999).
30. Spitzer, N. C. & Ribera, A. B. Development of electrical excitability in embryonic neurons: Mechanisms and roles. *J. Neurobiol.* **37**, 190–197 (1998).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

### Acknowledgements

We thank T. Chatila for the CaMK I and IV-dCT plasmids, and suggesting the use of KN93; M. Greenberg for the CaMK IV-dCTK75E plasmid; R. Maurer for the CaMK II-dCT plasmid; T. Soderling for the CaMK IV (HMDT→DED) plasmid; B. Howard for PC12 cells. We are grateful to R. Harris-Warrick, D. McCobb, K. Lynch, G. Chanfreau, P. Boutz, J. Underwood, K. Martin, M. Barad and H. Herschman for comments on the manuscript. This work is supported by the Howard Hughes Medical Institute.

Correspondence and requests for materials should be addressed to D.L.B. (e-mail: [dough@microbio.ucla.edu](mailto:dough@microbio.ucla.edu)). The genomic DNA sequences for mouse STREX and SloII87 have been deposited in GenBank under accession numbers AF295094 and AF295095, respectively.

# Neuropeptide Y functions as a neuroproliferative factor

D. E. Hansel\*, B. A. Eipper\*† & G. V. Ronnett\*‡

Departments of \*Neuroscience and ‡Neurology, The Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA

†Department of Physiology, University of Connecticut Health Science Center, Farmington, Connecticut 05405, USA

Neuropeptide Y (NPY) has a number of functions in mammalian physiology<sup>1–6</sup>. Here we identify a role for NPY in promoting proliferation of postnatal neuronal precursor cells. NPY is synthesized in the postnatal olfactory epithelium by sustentacular cells, previously proposed to function only in structural support<sup>7</sup>. Mice with a targeted deletion of NPY<sup>8</sup> contain half as many dividing olfactory neuronal precursor cells as do controls. Furthermore, NPY-deficient mice develop significantly fewer olfactory neurons by adulthood. NPY acts on multipotent neuronal precursor or basal cells to activate rapidly and transiently the extracellular signal-regulated kinase (ERK)1/2 subgroup of mitogen-activated protein kinases. The NPY Y1 receptor subtype appears to mediate this effect. The ability of NPY to induce neuronal precursor proliferation is mediated by protein kinase C (PKC), indicating an upstream PKC-dependent activation of ERK1/2. These results indicate that NPY may regulate neuronal precursor proliferation in the adult mammal.

NPY is a 36-amino-acid neuropeptide that is broadly expressed in the central and peripheral nervous system during development and adulthood<sup>1,2</sup>. It regulates feeding behaviour, gastrointestinal activity and central cardiovascular function, and influences seizure threshold and alcohol intake<sup>3–5</sup>. NPY messenger RNA is upregulated following peripheral axotomy and in pheochromocytoma and ganglioneuroblastoma tissue, although its role in neuronal proliferation is unknown<sup>6</sup>. To investigate the action of NPY in neuroproliferation, we examined the rodent olfactory epithelium. The olfactory epithelium is composed of immature and mature olfactory receptor neurons, non-neuronal sustentacular cells and proliferative basal cells, which comprise the multipotent progenitor cells of the olfactory epithelium<sup>7</sup>. The olfactory basal cells represent one of the few well-documented regenerating neuronal populations in the adult mammalian nervous system; the others are the stem cells of the rostral migratory stream, hippocampus and retina<sup>9,10</sup>.

We used immunohistochemical methods and radioimmunoassays to investigate NPY expression in the rat olfactory system. NPY was previously identified in the ensheathing cells of olfactory axon bundles of the adult rat<sup>11</sup>, but not within cells of the olfactory epithelium. We detected NPY expression in a subpopulation of neuronal cells in the olfactory epithelium at embryonic day 16. The NPY-expressing cells were identified as neurons by double-label immunofluorescence with neuron-specific tubulin (NST; Fig. 1a), a specific neuronal marker that is expressed early in the olfactory neuronal lineage in both immature neurons and dividing cells<sup>12,13</sup>. In the adult, however, NPY no longer co-localized with sites of NST expression, but was found in a subpopulation of sustentacular cells (Fig. 1b). Sustentacular cells comprise the outermost layer of olfactory epithelial cells, and are proposed to maintain epithelial structure and to secrete protective detoxifying enzymes<sup>14</sup>. This is the first report, to our knowledge, of a heterogeneous expression pattern within the sustentacular cell population.

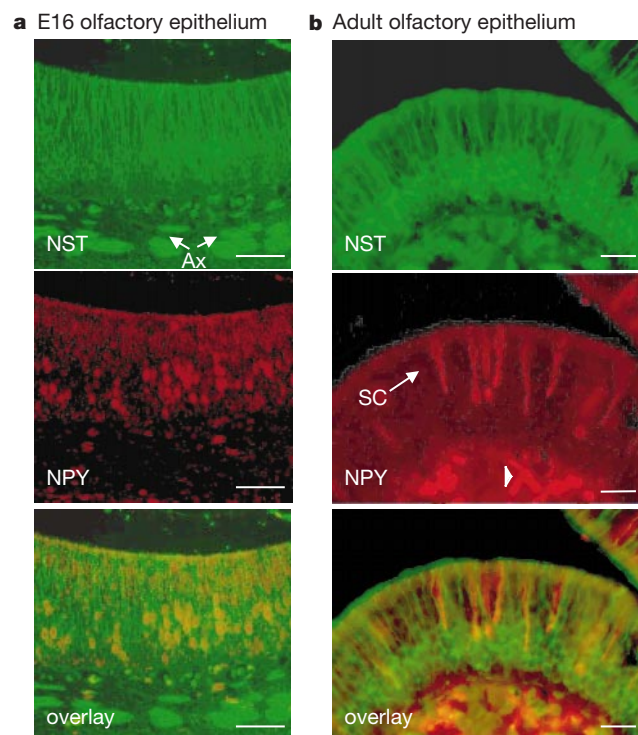
To confirm the existence of NPY in the olfactory epithelium, we quantified NPY content by radioimmunoassaying rat olfactory epithelium ( $0.56 \pm 0.7$  pmol NPY per mg protein (mean  $\pm$  s.e.m.)), olfactory bulb ( $1.95 \pm 0.28$  pmol NPY per mg), hypothalamus ( $6.17 \pm 1.12$  pmol NPY per mg) and primary

olfactory neuronal cultures (levels below the limit of detection ( $< 0.05$  pmol NPY per mg)). Although the arcuate nucleus of the hypothalamus contains the highest reported level of NPY in the central nervous system<sup>1</sup> the NPY content of the olfactory system is substantial in comparison to regions such as the cerebellum or pons<sup>1</sup>, which express very low levels of NPY.

We examined the adult olfactory epithelium of NPY-deficient mice generated by targeted gene deletion<sup>8</sup> and wild-type littermate controls to identify the potential role of NPY *in vivo*. Labelling of proliferating basal cells in NPY-deficient and control animals with 5-bromo-2'-deoxyuridine (BrdU) showed a roughly 50% decrease in the number of dividing cells in the NPY-deficient mice (Fig. 2a, b). The number of cells expressing Ki67, a marker of cycling cells that are not in G<sub>0</sub> (ref. 15), was also decreased in NPY-deficient mice, confirming a significant reduction in neuronal precursor proliferation with the loss of NPY (Fig. 2c).

To determine the overall consequences of NPY deficiency in the olfactory epithelium, we examined the neuronal populations of NPY-deficient and control mice. The number of neuron-specific tubulin<sup>12,13</sup> immunopositive cells was reduced in NPY-deficient mice (Fig. 2d). Neuronal cell counts were performed using O/E-1, a transcription factor present in cells of the olfactory neuronal lineage<sup>16</sup>; NPY-deficient mice contained around 25% fewer neurons than control mice (Fig. 2e, f). The difference between proliferation in the epithelium (50%) and total neuronal number (25%) in NPY-deficient and control mice may be secondary to compensatory changes in the olfactory epithelium. These results indicate that loss of NPY *in vivo* reduced neuronal precursor proliferation in the olfactory epithelium, as well as the overall number of neurons.

We used primary olfactory cultures<sup>17</sup> to examine the mechanism



**Figure 1** NPY expression in the olfactory epithelium. **a**, Expression of NPY in a subset of developing olfactory receptor neurons. All antibodies were used at a 1:1,000 dilution. Double-label immunofluorescence at embryonic day (E)16 shows NPY in cells positive for neuron-specific tubulin (NST). NPY is absent from the developing axon bundles (Ax). Scale bar, 50  $\mu$ m. **b**, Expression of NPY in non-neuronal sustentacular cells in the adult olfactory epithelium. NPY expression is detected in a subpopulation of sustentacular cells (SC) and in the underlying ensheathing cells (arrowhead). NPY in the adult was visualized using the NEN fluorescent tyramide system. Scale bar, 25  $\mu$ m.

of NPY action in olfactory receptor neurons. These cultures, which contain a predominance of basal cells and olfactory receptor neurons and a subpopulation of glial cells from the lamina propria, do not contain detectable levels of NPY by radioimmunoassay. Primary olfactory cultures contain very few sustentacular cells and, consequently, only about 2–3% of the cultured cells are immunopositive for NPY.

We verified NPY receptor binding at 4 °C and 37 °C in primary olfactory cultures (Fig. 3a). The NPY-binding cells were rounded and morphologically similar to the proliferative basal cells<sup>7</sup>; these cells accounted for around 1 out of 20 cells in the culture. Thus, NPY receptors are positioned to mediate an early step in olfactory neurogenesis.

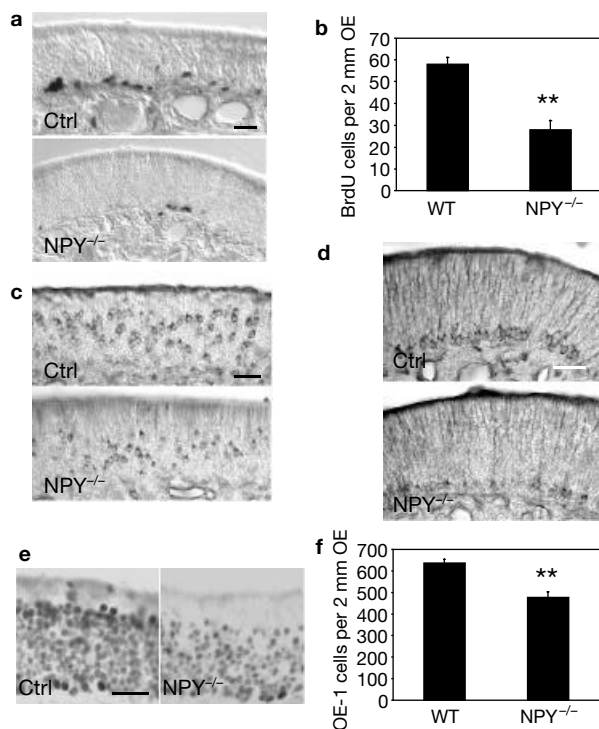
NPY can mediate the proliferation of vascular smooth muscle cells and colonic lamina propria lymphocytes<sup>18,19</sup>. To determine whether NPY could perform a similar function in the nervous system, we incubated primary olfactory cultures from 2-day-old rats with increasing concentrations ( $10^{-8}$  M to  $10^{-6}$  M) of NPY (Fig. 3b). NPY increased the number of neurons (NST-positive) in a dose-dependent manner, without influencing the glial cell (GFAP-positive) subpopulation. To determine whether this increase in

neuronal number was due to increased neuronal proliferation or to increased neuronal survival, we compared results from BrdU incorporation assays and 3' terminal uridine nick end labelling (TUNEL) apoptosis assays.

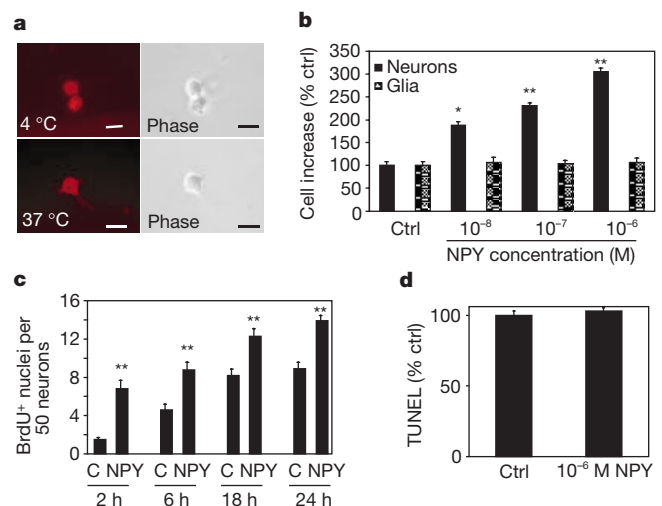
NPY increased the number of basal cells undergoing cell division to produce more neurons (BrdU; Fig. 3c). The increase in BrdU labelling between control and NPY-treated cells after only 2 h of exposure to NPY indicates that NPY may promote transition of basal cells into S phase of the cell cycle<sup>20</sup>. NPY did not alter the number of cells identified by TUNEL staining in the cultures (Fig. 3d), indicating that NPY does not affect the rate of apoptosis. These results indicate that NPY increases neuronal number by inducing neuroproliferation, identical to its role *in vivo*, without altering neuronal survival.

NPY can act through at least five receptors, including Y1, Y2, Y4, Y5 and Y6 (ref. 21). All of these receptors are G-protein-coupled and affect multiple, response-specific second messenger cascades. Evidence from studies performed in transfected cells indicates that NPY can signal through various intracellular pathways, including calcium and cyclic AMP<sup>22</sup>. The Y1 and Y2 receptors are the most prevalent NPY receptors in the mammalian brain, and subtype agonists were used to determine which receptor promotes NPY-mediated neuroproliferation<sup>23</sup>. Incubation of cells with the Y1/Y5 agonist [Leu<sup>31</sup>, Pro<sup>34</sup>]NPY mimicked the effect of NPY on neuroproliferation (Fig. 4a); the diminished effect of the higher dose may reflect receptor desensitization. The Y2 agonist NPY(3-36) did not affect neuronal number.

To confirm that NPY acts through the Y1 receptor, rather than the Y5 receptor, to induce neuronal precursor or basal cell proliferation, we incubated primary olfactory cultures with  $10^{-6}$  M of the Y1-



**Figure 2** NPY-deficient mice<sup>8</sup> show decreased precursor proliferation and decreased neuronal number. **a**, BrdU labelling of adult mice shows decreased postnatal neurogenesis in the precursor cell population of NPY<sup>-/-</sup> mice. **b**, NPY-deficient mice have fewer dividing precursors. BrdU labelling of NPY<sup>-/-</sup> mice and wild-type littermate controls for 24 h shows 58 BrdU-positive cells per 2 mm of olfactory epithelium (OE) in wild-type mice, as compared to 29 BrdU-positive cells per 2 mm of olfactory epithelium in NPY<sup>-/-</sup> mice; six wild-type and five null mice were compared. Double asterisk,  $P < 0.001$  relative to control, using student's two-tailed  $t$ -test with samples of unequal variance. **c**, Ki67 immunostaining of NPY<sup>-/-</sup> mice and control mice. **d**, NST immunostaining of NPY<sup>-/-</sup> mice and control mice suggesting decreased numbers of immunoreactive cells in NPY<sup>-/-</sup> mice. **e**, O/E-1 immunostaining of NPY<sup>-/-</sup> and control mice shows fewer post-mitotic neurons throughout the NPY<sup>-/-</sup> olfactory epithelium. **f**, Quantification of O/E-1 immunopositive cells comparing NPY<sup>-/-</sup> and control mice, examining 6–7 comparable regions of olfactory epithelium. Control animals contained  $633 \pm 19$  cells per 2 mm epithelium, whereas NPY<sup>-/-</sup> mice contained  $470 \pm 26$  cells per 2 mm epithelium. Double asterisk,  $P < 0.001$  using the student's two-tailed  $t$ -test with samples of unequal variance. Scale bars, 25  $\mu$ m.



**Figure 3** NPY stimulates neuroproliferation. **a**, Fluorescein-tagged NPY (30 nM; NEN) binds to the surface of a population of cells in primary neuronal cultures at 4 °C and 37 °C. Blocking control was performed by co-incubating fluorescently tagged NPY with 3  $\mu$ M unlabelled peptide. Images are pseudocoloured red. Scale bar, 20  $\mu$ m. **b**, Twenty-four hour incubation of 1-day-old primary olfactory neuronal cultures with NPY (Peninsula Lab) increases neuronal number, as defined by NST staining. NPY does not alter glial cell number, as defined by glial fibrillary acidic protein staining. Cell counts are reported as percentage of stained cells per high power field relative to control. 100% NST is 40 neurons per field. 100% GFAP is 25 glial cells per field. Incubation with or without NGF does not alter neuronal or glial number, or responsiveness to NPY. Error bars, s.e.m. Asterisk,  $P < 0.05$ ; double asterisk,  $P < 0.001$ , relative to control (student's two-tailed  $t$ -test with samples of unequal variance). **c**, NPY promotes neuronal proliferation. Incubation of cells with  $10^{-6}$  M NPY for the time indicated increased BrdU<sup>+</sup> nuclei per 50 NST<sup>+</sup> cells (neurons) at all time points.  $P$ -values given relative to control (C) values at each time point. **d**, TUNEL staining of primary olfactory cultures incubated with or without  $10^{-6}$  M NPY indicates no significant change in the rate of apoptosis.



receptor-specific antagonist BIBP3226. Incubation of cells with NPY and BIBP3226 abolished NPY-induced neuroproliferation (Fig. 4b), indicating that the actions of NPY were probably mediated by the Y1 receptor; BIBP3226 alone did not alter neuronal numbers. These results are consistent with a role for the NPY Y1 receptor subtype in neuronal precursor proliferation, although an unknown NPY receptor, functioning in a similar manner, cannot be ruled out.

To determine whether the ability of NPY to increase neuronal precursor proliferation was a direct or an indirect effect on basal cells, we localized the NPY receptor. The proliferative neuronal precursor/basal cells in the olfactory epithelium rest along the basement membrane and undergo proliferation into adulthood to generate new neurons<sup>7</sup>. *In vitro* tissue binding with 50 nM of fluorescently labelled NPY identified binding of the peptide to most of the basally situated cells (Fig. 4c, d). Thus, the NPY receptor is localized to the appropriate cells to mediate directly the neuroproliferative response demonstrated in primary culture. Competition with  $10^{-6}$  M unlabelled NPY blocked the ability of the labelled peptide to bind to these cells (Fig. 4e).

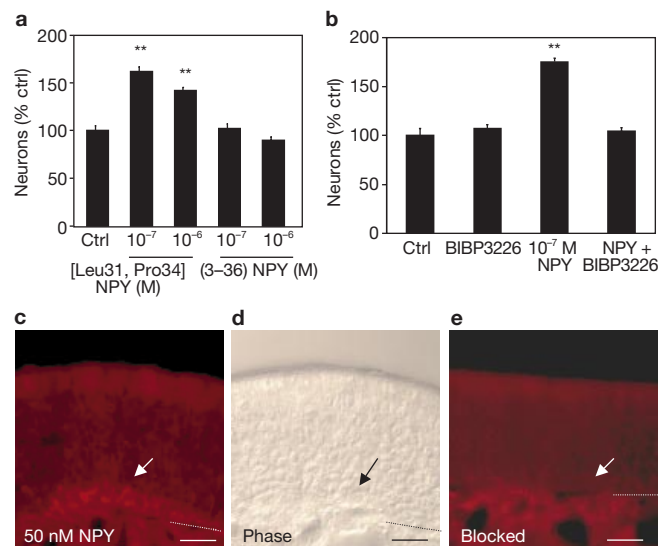
In other systems, the ERK subgroup of the mitogen-activated protein (MAP) kinases is important in the regulation of cellular proliferation<sup>24,25</sup>. NPY increases the phosphorylation of ERK1/2 in cardiomyocytes and in a variety of clonal cell lines, although this effect has not been described in neurons<sup>26–28</sup>. ERK1/2 is activated by dual phosphorylation on threonine and tyrosine residues by MAP or ERK kinases (MEK)1/2 (refs 24, 25). We therefore examined the role of ERK1/2 in mediating NPY-induced neuronal proliferation.

Incubation of cells with 100  $\mu$ M PD98059, a potent MEK inhibitor, blocked the NPY-mediated increase in neuronal number (Fig. 5a). PD98059 alone did not alter neuronal number. Incubation of cells with genistein, a general tyrosine kinase inhibitor

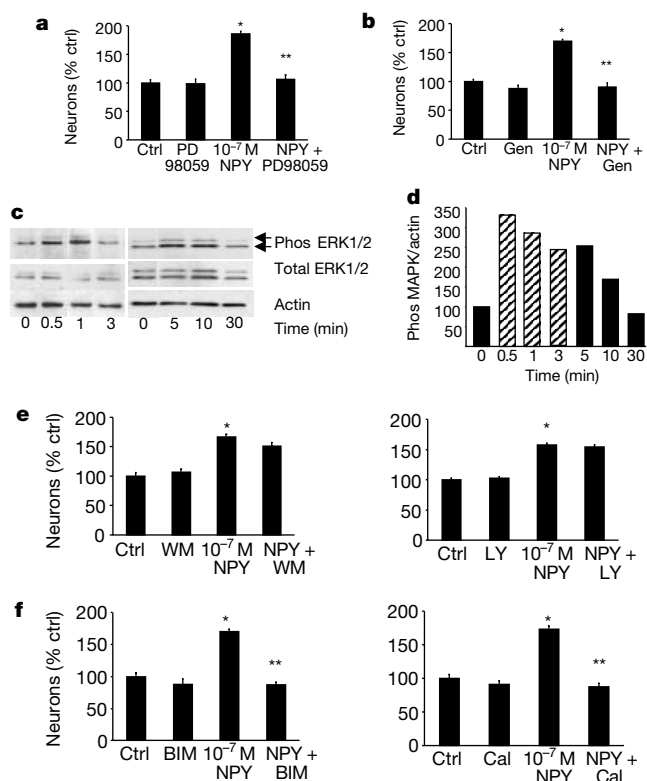
that blocked the activation of ERK1/2, also prevented the NPY-induced increase in neuronal number (Fig. 5b). As rapid and transient phosphorylation of ERK1/2 can lead to long-term alterations in cellular proliferation<sup>24,25</sup>, we examined phosphorylation of ERK1/2 after NPY addition for times up to 30 min (Fig. 5c, d). NPY-induced phosphorylation of ERK1/2 was detected 30 s after treatment, with a sustained elevation in phosphorylation for up to 10 min, and a return to baseline levels of phosphorylation by 30 min. The maximum response was observed at 30 s, with phosphorylation increased threefold over baseline (Fig. 5d).

The ERK1/2 pathway is regulated by a variety of kinases, including phosphatidylinositol-3-OH kinase (PI(3)K), PKC and protein kinase A<sup>24,25</sup>. NPY has been proposed to stimulate the phosphorylation of ERK1/2 in cardiomyocytes through a PI(3)K-dependent mechanism<sup>26–28</sup>. To determine whether NPY causes neuroproliferation through activation of PI(3)K or PKC<sup>29</sup>, we incubated cells with inhibitors specific for each pathway.

Incubation of cells with either 100 nM wortmannin or 20  $\mu$ M



**Figure 4** Y1 receptors mediate the NPY response. **a**, Incubation of cultures for 24 h with the Y1/Y5 agonist [Leu<sup>31</sup>, Pro<sup>34</sup>]NPY (Peninsula) increases the number of NST<sup>+</sup> cells. Incubation with the Y2 agonist NPY(3–36) (Peninsula) yielded no response. *P*-values given relative to control (ctrl) values. **b**, Twenty-four hour co-incubation with the Y1 receptor antagonist BIBP3226 (1  $\mu$ M) and NPY blocked the ability of NPY to increase neuronal number. BIBP3226 alone had no effect. **c**, NPY binds to neuronal precursor cells. Incubation of 50 nM fluorescently tagged NPY (NEN) with Bouin's fixed rat olfactory epithelium for 2 h showed binding of NPY to the basal cell layer (arrow). Dotted line, separation between olfactory epithelium and underlying lamina propria. Scale bar, 25  $\mu$ m. **d**, Phase microscopy view of **c**; arrow indicates location of stained cells (basal cell layer). **e**, Co-incubation with 50 nM fluorescent NPY and 5  $\mu$ M unlabelled peptide (Peninsula) abolished binding to most of the basal cells (arrow). Nonspecific binding in the underlying lamina propria was evident in both bound and control conditions.



**Figure 5** NPY induces neuroproliferation through ERK1/2 phosphorylation and activation.

**a**, MEK blockade abolishes NPY-induced neuroproliferation. Twenty-four hour incubation of cells with NPY increases neuronal number. Incubation with PD98059 (100  $\mu$ M; Calbiochem), an inhibitor of MEK activation, does not alter NST<sup>+</sup> cells, but co-incubation with PD98059 and NPY blocks the ability of NPY to induce neuronal proliferation. Asterisk, *P* < 0.001 between NPY and control. Double asterisk, *P* < 0.001 between NPY and NPY plus PD98059. **b**, Tyrosine kinase inhibition blocks the NPY-induced increase in neuronal number. Twenty-four hour co-incubation of cells with NPY and 100  $\mu$ M genistein (Gen; Calbiochem, concentration within manufacturer's recommended range) prevented the NPY effect. **c**, NPY promotes rapid phosphorylation of ERK1/2. **d**, Western blots from four experiments were densitized using the Scion Image Program. Ratios represent the increase over control (0 min) levels of phosphorylated ERK1/2 normalized to levels of actin. **e**, NPY does not induce neuroproliferation through the PI(3)K pathway. Co-incubation of cultures for 24 h with NPY and either 100 nM wortmannin (WM; Calbiochem) or 20  $\mu$ M LY294002 (LY; Calbiochem) did not alter the NPY-induced increase in neuronal number. **f**, PKC mediates the NPY effect. Co-incubation of cells for 24 h with NPY and either 10  $\mu$ M BIM-1 (Calbiochem) or 0.5  $\mu$ M calphostin (Cal; Calbiochem) blocked the ability of NPY to induce neuroproliferation. Asterisk, *P* < 0.001 between NPY and control. Double asterisk, *P* < 0.001 between NPY and NPY plus drug.

LY294002, which specifically inhibit PI(3)K, did not block the ability of NPY to increase neuronal number (Fig. 5f). However, incubation of cells with the PKC inhibitors bisindolylmaleimide I (BIM-1, 10  $\mu$ M) or calphostin (Cal, 0.5  $\mu$ M) abolished the effect of NPY on neuroproliferation (Fig. 5g). These results indicate that NPY regulates upstream activation of ERK1/2 by a PKC-dependent pathway.

This is the first report, to our knowledge, of NPY-induced neuroproliferation. NPY, expressed by a subset of non-neuronal sustentacular cells in the adult olfactory epithelium, acts on neuronal precursor cells (basal cells) to induce cell-cycle activation and division. NPY acts through the Y1 receptor to increase neuronal precursor proliferation; this effect is mediated downstream through a kinase cascade involving PKC and ERK1/2. NPY can induce cardiomyocyte hypertrophy by a similar mechanism, although the upstream regulator of ERK1/2 activation in cardiomyocytes involves a PI(3)K-dependent pathway<sup>26–28</sup>.

Although the olfactory system has been considered unusual in its ability to maintain multipotent stem-cell-like populations, the recent discovery of stem cells in the retina, hippocampus and rostral migratory stream<sup>9,10</sup> indicates that postnatal neurogenesis could be a potentially widespread mechanism by which the nervous system repopulates certain neuronal lineages. NPY is broadly expressed in the developing nervous system and is maintained at moderately high levels in the adult, although its actions vary according to the target cell<sup>1,2</sup>. The ability of NPY to stimulate neuroproliferation of the olfactory neuronal precursor cell could imply a similar function on central nervous system stem cells, and may be useful for fields such as stem cell therapy. □

## Methods

### Antibodies and immunofluorescent immunostaining

All experimental protocols were approved by the Johns Hopkins University Institutional Animal Care and Use Committee, and all applicable guidelines from the *National Institutes of Health Guide for the Care and Use of Laboratory Animals* were followed. Five-week-old Sprague–Dawley rats (Harlan) were anaesthetized with Xylaket and perfused with phosphate-buffered saline (PBS) followed by Bouin's fixative. Tissue was dissected, post-fixed overnight in Bouin's fixative at 4°C, washed in PBS, placed in 20% sucrose and embedded in Tissue-Tek (Sakura Finetek). Paraffin-embedded paraformaldehyde fixed embryos were obtained from Novagen. Rabbit polyclonal NPY antibody JH3 (5 ml) was affinity-purified using NPY peptide (1–36)NH<sub>2</sub> (1 mg; Peninsula) linked to Affigel 10 resin (2 ml; Bio-Rad). For double-label immunofluorescence, sections were washed in PBS, permeabilized with 0.05% SDS in PBS, incubated with 1% H<sub>2</sub>O<sub>2</sub>, blocked in TNB buffer (NEN) and then incubated overnight at 4°C with monoclonal antibody against NST at 1:1,000 (BabCo). The following day, sections were incubated with affinity purified NPY antibody at 1:1,000 using the tyramide-direct protocol (NEN).

### DAB immunostaining

Four-week-old male NPY-deficient<sup>8</sup> and wild-type littermate controls were used. Tissue was dissected and sectioned as described. Sections were hydrated in PBS, permeabilized in 0.05% SDS in PBS, blocked in appropriate 4% normal serum with 1% BSA (Sigma) and incubated overnight at 4°C with NST monoclonal antibody at 1:1,000 (BabCo), Ki67 monoclonal antibody at 1:100 (Immunotech) and O/E-1 polyclonal antibody at 1:100 (a gift from the laboratory of R. Reed). The following day, sections were incubated in 1% H<sub>2</sub>O<sub>2</sub> and developed according to the ABC protocol (Vector Laboratories). Six or seven fields were counted per animal for O/E-1.

### Radioimmunoassay

NPY levels were quantified by radioimmunoassay<sup>30</sup> in acetic acid extracts of olfactory epithelium, olfactory bulb, hypothalamus and primary olfactory neuron cultures. Radioimmunoassay was performed using [<sup>125</sup>I]-labelled NPY (NEN) and anti-NPY JH3 antibody diluted 1:40,000.

### BrdU labelling

One-day-old cultures were incubated with BrdU reagent (Boehringer Mannheim) according to instructions with or without NPY. After incubation, cells were rinsed in PBS, fixed in ice-cold 70% EtOH in 50 mM glycine, pH 2.0, and processed for BrdU staining. Cells were then washed in PBS, blocked with 4% normal horse serum (Vector Laboratories), incubated overnight at 4°C with NST antibody 1:1,000, rinsed in PBS and processed with the Elite ABC Kit for DAB staining (Vector). Cells were counted per high power field using 9–10 fields per well. For animal experiments, mice were injected with 50 mg kg<sup>-1</sup> BrdU labelling reagent, killed 24 h after injection and perfused with Bouin's fixative.

### TUNEL staining

One-day-old cultures were treated with or without peptide for 48 h, rinsed in PBS and fixed in MeOH:acetone (1:1). Cells were then treated with 3% H<sub>2</sub>O<sub>2</sub>, incubated in TdT buffer (33 mM Tris, pH 7.2, 140 mM NaCacodylate, 1 mM CoCl<sub>2</sub>), treated with 0.12 U  $\mu$ l<sup>-1</sup> terminal transferase (Sigma) and a 1:200 dilution of biotin-16-2'-deoxyuridine-5'-triphosphate (Boehringer Mannheim), washed in  $\times$ 2 SSC, washed in PBS, blocked with 3% normal horse serum (Vector) and developed using the Elite ABC Kit (Vector).

### Cell culture

We used around fifty 0–1-day-old rat pups per experiment for culture. Cells were grown for one day after plating in 12-well plates. After incubation with 10<sup>-6</sup> M NPY for the appropriate time, cells were lysed with boiling sample buffer (50 mM Tris-HCl, pH 6.8, 2% SDS, 1%  $\beta$ -mercaptoethanol, 10% glycerol, 0.1% bromophenol blue) containing 0.3 mg ml<sup>-1</sup> PMSF, 50  $\mu$ g ml<sup>-1</sup> lima bean trypsin inhibitor, 2  $\mu$ g ml<sup>-1</sup> leupeptidin, 16  $\mu$ g ml<sup>-1</sup> benzamide, 2  $\mu$ g ml<sup>-1</sup> pepstatin, 5 mM EGTA, 5 mM Na<sub>2</sub>EDTA, 1 mM sodium orthovanadate, 10 mM sodium phosphate and 50 mM sodium fluoride. Incubation media included 0.5% dialysed fetal bovine serum (Gibco BRL) and 2.5 ng NGF per ml medium. Samples were sonicated for 30 s, boiled for 5 min, cooled on ice and spun at 14,000 r.p.m. at 4°C for 10 min, aliquoted, and stored at –80°C.

### Western blotting

Olfactory cultures were incubated with 10<sup>-6</sup> M NPY for times from 30 s to 30 min, then harvested for western blotting. Blots were probed using an antibody specific for phosphorylated ERK1/2 (1:1,000; NEB). Standardization was performed by re-probing the blots with antibody for ERK1/2 (1:1,000; NEB) and actin (1:1,000; Sigma).

Received 24 November 2000; accepted 19 February 2001.

- Allen, Y. S. *et al.* Neuropeptide Y distribution in the rat brain. *Science* **221**, 877–879 (1983).
- Danger, J. M. *et al.* Neuropeptide Y: localization in the central nervous system and neuroendocrine functions. *Fundam. Clin. Pharmacol.* **4**, 307–340 (1990).
- Baraban, S. C., Hollopeter, G., Erickson, J. C., Schwartzkroin, P. A. & Palmiter, R. D. Knock-out mice reveal a critical antiepileptic role for neuropeptide Y. *J. Neurosci.* **17**, 8927–8936 (1997).
- Hokfelt, T. *et al.* Neuropeptide Y: some viewpoints on a multifaceted peptide in the normal and diseased nervous system. *Brain Res. Rev.* **26**, 154–166 (1998).
- Thiele, T. E., Marsh, D. J., Ste. Marie, L., Bernstein, L. L. & Palmiter, R. D. Ethanol consumption and resistance are inversely related to neuropeptide Y levels. *Nature* **396**, 366–369 (1998).
- Adrian, T. E. *et al.* Neuropeptide Y in pheochromocytomas and ganglioneuroblastomas. *Lancet* **2**, 540–542 (1983).
- Huard, J. M., Youngentob, S. L., Goldstein, B. J., Luskin, M. B. & Schwob, J. E. Adult olfactory epithelium contains multipotent progenitors that give rise to neurons and non-neuronal cells. *J. Comp. Neurol.* **400**, 469–486 (1998).
- Erickson, J. C., Clegg, K. E. & Palmiter, R. D. Sensitivity to leptin and susceptibility to seizures of mice lacking neuropeptide Y. *Nature* **381**, 415–421 (1996).
- Gage, F. H. Mammalian neural stem cells. *Science* **287**, 1433–1438 (2000).
- Tropepe, V. *et al.* Retinal stem cells in the adult mammalian eye. *Science* **287**, 2032–2036 (2000).
- Ubink, R., Halasz, N., Zhang, X., Dagerlind, A. & Hokfelt, T. Neuropeptide tyrosine is expressed in ensheathing cells around the olfactory nerves in the rat olfactory bulb. *Neuroscience* **60**, 709–726 (1994).
- Roskams, A. J. L., Cai, X. & Ronnett, G. V. Expression of neuron-specific beta-III tubulin during olfactory neurogenesis in the embryonic and adult rat. *Neuroscience* **83**, 191–200 (1998).
- Memberg, S. P. & Hall, A. K. Dividing neuron precursors express neuron-specific tubulin. *J. Neurobiol.* **27**, 26–43 (1995).
- Morrison, E. E. & Moran, D. T. *Handbook of Olfaction and Gustation* (ed. Doty, R. L.) 79–81 (Marcel Dekker, New York, 1995).
- Rijzewijk, J. J., Van Erp, P. E. & Bauer, F. W. Two binding sites for Ki67 related to quiescent and cycling cells in human epidermis. *Acta. Dermatol. Venerol.* **69**, 512–515 (1989).
- Davis, J. A. & Reed, R. R. Role of Olf-1 and Pax-6 transcription factors in neurodevelopment. *J. Neurosci.* **16**, 5082–5094 (1996).
- Ronnett, G. V., Hester, L. D. & Snyder, S. H. Primary culture of neonatal rat olfactory neurons. *J. Neurosci.* **11**, 1243–1255 (1991).
- Shigeri, Y. & Fujimoto, M. Neuropeptide Y stimulates DNA synthesis in vascular smooth muscle cells. *Neurosci. Lett.* **149**, 19–22 (1993).
- Elitsur, Y. *et al.* Neuropeptide Y (NPY) enhances proliferation of human colonic lamina propria lymphocytes. *Neuropeptides* **26**, 289–295 (1994).
- Nowakowski, R. S., Lewin, S. B. & Miller, M. W. Bromodeoxyuridine immunohistochemical determination of the lengths of the cell cycle and the DNA-synthetic phase for an anatomically defined population. *J. Neurocytol.* **18**, 311–318 (1989).
- Michel, M. C. *et al.* XVI. International Union of Pharmacology recommendations for the nomenclature of neuropeptide Y, peptide YY, and pancreatic polypeptide receptors. *Pharmacol. Rev.* **50**, 143–150 (1998).
- Balasubramaniam, A. A. Neuropeptide Y family of hormones: receptor subtypes and antagonists. *Peptides* **18**, 445–457 (1997).
- Naveilhan, P., Neveu, I., Arenas, E. & Ernfor, P. Complementary and overlapping nervous system. *Neuroscience* **87**, 289–302 (1998).
- Lopez-Illasaca, M. Signaling from G-protein-coupled receptors to mitogen-activated protein (MAP)-kinase cascades. *Biochem. Pharmacol.* **56**, 269–277 (1998).
- English, J. *et al.* New insights into the control of MAP kinase pathways. *Exp. Cell Res.* **253**, 255–270 (1999).
- Goldberg, Y., Taimor, G., Piper, H. M. & Schluter, K. D. Intracellular signaling leads to the hypertrophic effect of neuropeptide Y. *Am. J. Physiol.* **275**, C1207–C1215 (1998).

27. Keffel, S., Schmidt, M., Bischoff, A. & Michel, M. C. Neuropeptide-Y stimulation of extracellular signal-regulated kinases in human erythroleukemia cells. *J. Pharmacol. Exp. Ther.* **291**, 1172–1178 (1999).
28. Pellioux, C. *et al.* Neuropeptide Y (NPY) potentiates phenylephrine-induced mitogen-activated protein kinase activation in primary cardiomyocytes via NPY Y5 receptors. *Proc. Natl Acad. Sci. USA* **97**, 1595–1600 (2000).
29. Mannon, P. J. & Raymond, J. R. The neuropeptide Y/peptide YY Y1 receptor is coupled to MAP kinase via PKC and Ras in CHO cells. *Biochem. Biophys. Res. Commun.* **246**, 91–94 (1998).
30. Dickerson, I. M. & Mains, R. E. Cell-type specific posttranslational processing of peptides by different pituitary cell lines. *Endocrinology* **127**, 133–140 (1990).

## Acknowledgements

We thank A. Hanlon and M. Bell for technical support, and R. Mains and C. Hansel for comments on the manuscript. The NPY-deficient mice were a gift from R. D. Palmiter at the University of Washington; mouse genotyping was performed by G. Miura in the laboratory of R. D. Palmiter.

Correspondence and requests for materials should be addressed to G.V.R. (e-mail: gronnett@bs.jhmi.edu).

# Insulin-stimulated GLUT4 translocation requires the CAP-dependent activation of TC10

Shian-Huey Chiang<sup>†‡</sup>, Christian A. Baumann<sup>†‡</sup>, Makoto Kanzaki<sup>§</sup>, Debbie C. Thurmond<sup>§</sup>, Robert T. Watson<sup>§</sup>, Cheryl L. Neudauer<sup>||</sup>, Ian G. Macara<sup>||</sup>, Jeffrey E. Pessin<sup>§</sup> & Alan R. Saltiel<sup>†‡</sup>

<sup>\*</sup> Cellular and Molecular Biology Graduate Program, University of Michigan,

<sup>‡</sup> Department of Medicine and Physiology, University of Michigan School of Medicine, Ann Arbor, Michigan 48109, USA

<sup>†</sup> Department of Cell Biology, Pfizer Global Research and Development, Ann Arbor, Michigan 48105, USA

<sup>§</sup> Department of Physiology & Biophysics, The University of Iowa, Iowa City, Iowa 52242, USA

<sup>||</sup> Center for Cell Signaling, University of Virginia, Charlottesville, Virginia 22908, USA

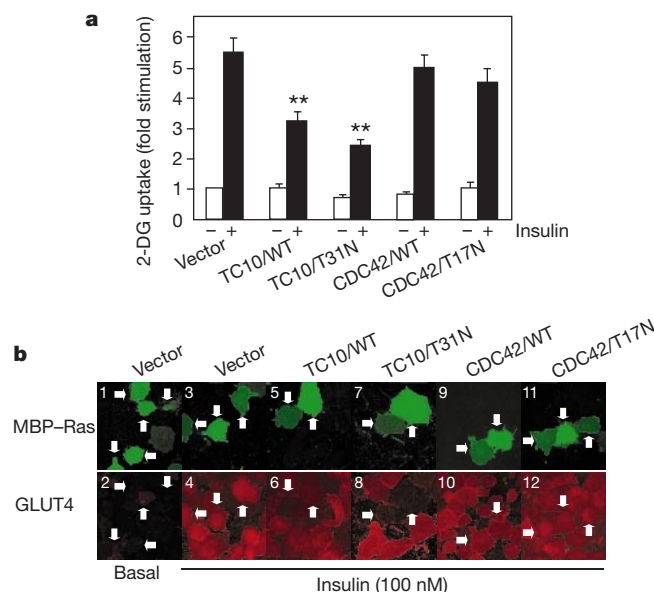
The stimulation of glucose uptake by insulin in muscle and adipose tissue requires translocation of the GLUT4 glucose transporter protein from intracellular storage sites to the cell surface<sup>1–6</sup>. Although the cellular dynamics of GLUT4 vesicle trafficking are well described, the signalling pathways that link the insulin receptor to GLUT4 translocation remain poorly understood. Activation of phosphatidylinositol-3-OH kinase (PI(3)K) is required for this trafficking event, but it is not sufficient to produce GLUT4 translocation<sup>7</sup>. We previously described a pathway involving the insulin-stimulated tyrosine phosphorylation of Cbl, which is recruited to the insulin receptor by the adapter protein CAP<sup>8,9</sup>. On phosphorylation, Cbl is translocated to lipid rafts. Blocking this step completely inhibits the stimulation of GLUT4 translocation by insulin<sup>10</sup>. Here we show that phosphorylated Cbl recruits the CrkII–C3G complex to lipid rafts, where C3G specifically activates the small GTP-binding protein TC10. This process is independent of PI(3)K, but requires the translocation of Cbl, Crk and C3G to the lipid raft. The activation of TC10 is essential for insulin-stimulated glucose uptake and GLUT4 translocation. The TC10 pathway functions in parallel with PI(3)K to stimulate fully GLUT4 translocation in response to insulin.

Many receptor tyrosine kinases use small GTP-binding proteins as molecular switches to convert proximal tyrosine phosphorylation into the activation of serine/threonine kinase cascades<sup>11</sup>. In addition, small GTP-binding proteins are critical for intracellular

vesicle trafficking<sup>12</sup>. We have examined the modulation of insulin-stimulated glucose uptake and GLUT4 translocation by several small GTP-binding proteins<sup>1</sup>. Expression of wild-type, dominant-interfering and constitutively active mutants of Ras, Rap1, Rac1, TC21, RhoA and RhoD in 3T3L1 adipocytes had no discernible effect on this biological response (data not shown). In contrast, expression of wild-type TC10 (TC10/WT) and a dominant-interfering TC10 mutant (TC10/T31N) significantly inhibited insulin-stimulated glucose uptake (Fig. 1a).

As the transfection efficiency was about 50%, it is likely that expression of the TC10/T31N mutant almost completely inhibited glucose uptake in the transfected cell population. Out of the known GTP-binding proteins, Cdc42 has the greatest degree of sequence similarity to TC10, with an overall sequence identity of 69% and similarity of 83%. However, expression of wild-type Cdc42 (Cdc42/WT) or the dominant-interfering Cdc42 mutant (Cdc42/T17N) had no statistically significant effect on insulin-stimulated glucose uptake (Fig. 1a).

To determine whether TC10 inhibits insulin-stimulated glucose uptake through a blockade of GLUT4 translocation, we microinjected 3T3L1 adipocytes along with expression plasmids for TC10 or Cdc42 (Fig. 1b). The cells were microinjected with the carboxy-terminal domain of Ras fused to the maltose-binding protein (MBP–Ras) as a marker for plasma-membrane in injected cells<sup>2–4</sup>. As GLUT4 is localized predominantly in intracellular vesicles in the basal state, there is only a low level of GLUT4 immunofluorescence in the plasma-membrane sheets derived from both the microinjected and non-microinjected cells. As typically observed, insulin produced the robust appearance of plasma-membrane GLUT4 immunofluorescence in both the microinjected and surrounding non-injected cells. Microinjection of either TC10/WT or TC10/T31N inhibited insulin-stimulated translocation of GLUT4, in



**Figure 1** Expression of TC10 inhibits insulin-stimulated glucose uptake and GLUT4 translocation in 3T3L1 adipocytes. **a**, Differentiated 3T3L1 adipocytes were transfected with TC10 and Cdc42 mutants. The cells were left untreated or were stimulated with 100 nM insulin for 30 min. The rate of [<sup>3</sup>H]2-deoxyglucose (2-DG) uptake was determined. Results are the mean  $\pm$  s.e. of 3–5 independent experiments from individual experiments performed in triplicate with a transfection efficiency of  $\sim$ 50%. Two asterisks,  $P < 0.01$ . **b**, Differentiated 3T3L1 adipocyte nuclei were microinjected with 0.2 mg ml<sup>-1</sup> of MBP–Ras plus either the empty vector or the cDNAs encoding TC10/WT, TC10/T31N, CDC42/WT or CDC42/T17N. The cells were allowed to recover for 24 h, and left untreated or treated with 100 nM insulin for 30 min.



comparison to the surrounding non-injected cells. Microinjection of Cdc42/WT or Cdc42/T17N had no significant effect on insulin-stimulated GLUT4 translocation. We scored 30–60 plasma-membrane sheets per condition from 4 independent experiments, which showed that the complementary DNAs for TC10/WT, TC10/T31N, Cdc42/WT and Cdc42/T17N inhibited insulin-stimulated GLUT4 translocation in  $66 \pm 8$ ,  $75 \pm 8$ ,  $7 \pm 3$  and  $8 \pm 4\%$  of the microinjected 3T3L1 adipocytes, respectively.

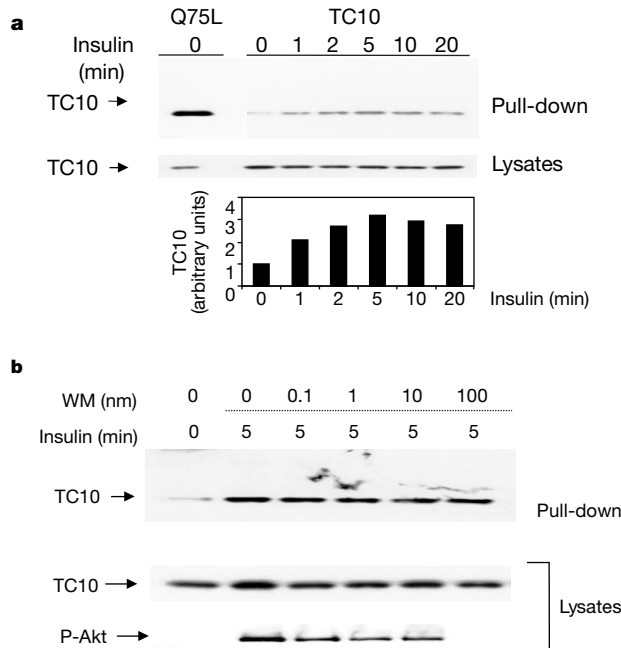
As a control for GLUT4 translocation specificity, we also examined the effect of TC10 and Cdc42 expression on insulin-stimulated GLUT1 translocation (Supplementary Information Fig. 1). Although insulin has a more modest effect on the translocation of GLUT1 as compared with GLUT4 (refs 5, 6), microinjection of both the TC10 and Cdc42 cDNAs had no effect.

To confirm the effect of TC10 on GLUT4 translocation, we co-expressed TC10 with a GLUT4-enhanced green fluorescent protein (EGFP) fusion construct in 3T3L1 adipocytes (Supplementary Information Fig. 2). Co-expression of GLUT4-EGFP with TC10/WT or TC10/T31N markedly decreased the number of the cells responding to insulin. In contrast, overexpression of Cdc42/WT or Cdc42/T17N had no effect on insulin-stimulated GLUT4-EGFP translocation. In addition, expressing the constitutively active mutant TC10/Q75L and GTP-cycling mutant TC10/F42L resulted in a substantial inhibition of insulin-stimulated GLUT4 translocation. As a measure of GLUT4 vesicle specificity, expression of TC10 or Cdc42 had no significant effect on the insulin-stimulated translocation of the cation-independent mannose-6-phosphate receptor (data not shown).

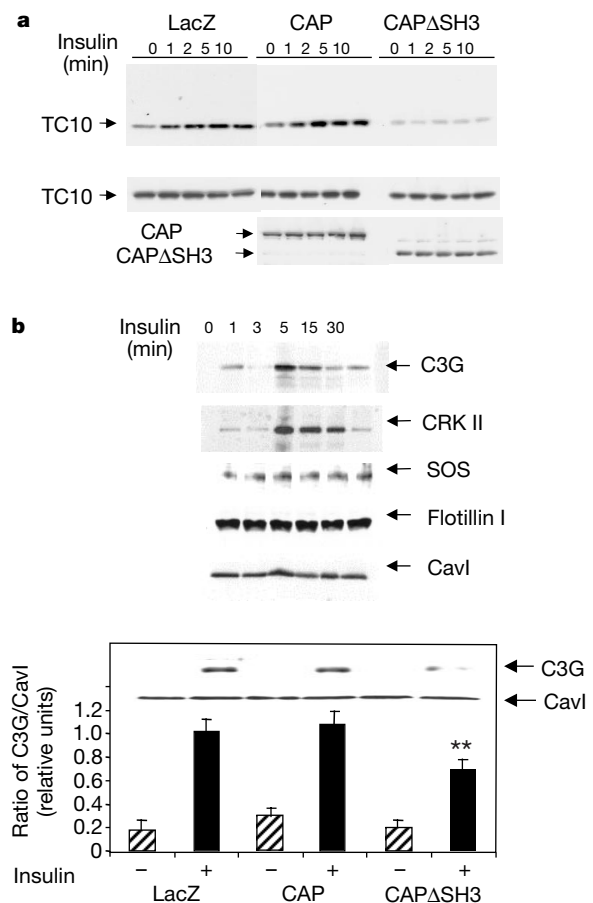
The inhibition of insulin-stimulated glucose uptake and GLUT4 translocation through overexpression of the TC10/T31N dominant-

interfering mutant is consistent with its role as a signalling intermediate; however, it was surprising that overexpression of the wild-type TC10 protein also inhibited this biological response. TC10 has a lower intrinsic GTPase activity than has Cdc42 (ref. 13). If a GTPase-activating activity limits the process, overexpression of TC10 would cause increased basal TC10-GTP, and thus behave like a constitutively active mutant. Indeed, both constitutively active and dominant-negative rhoA or Rac1 proteins perturbed tight junction gate function<sup>14</sup>. Thus, the overexpressed mutant GTPases might displace cycling endogenous GTPases from targets. As mutant proteins remain bound to the effectors, overexpression results in quick domination of the mutant phenotype. Overexpressing wild-type TC10 may alter binding stoichiometry and/or mis-target the effectors to wrong locations and thereby prevent the appropriate propagation of the signal.

Several studies have shown that in the active (GTP-bound) con-



**Figure 2** Insulin activates TC10 through a PI(3)K-independent pathway in 3T3L1 adipocytes. **a**, Fully differentiated 3T3L1 adipocytes were electroporated with 60  $\mu$ g of a cDNA encoding HA-tagged TC10. Cells were starved for 3 h before treatment with 100 nM insulin. Bottom, lysates from different time points were immunoblotted with anti-HA antibody. Top, lysates were incubated with GST-Pak1 for 1 h, followed by 3 washes and immunoblotted with anti-HA (top). **b**, 3T3L1 adipocytes were electroporated with HA-tagged TC10, incubated for 30 min with increasing doses of wortmannin (WM; 0–100 nM), and treated with insulin for 5 min. Top, GST-Pak1 pull-down assay probed with anti-HA. Middle, anti-HA immunoblots with lysates from different treatments. Bottom, phospho-Akt blot with the same lysates.



**Figure 3** TC10 is activated by insulin through a CAP-dependent pathway. **a**, 3T3L1 adipocytes were co-electroporated with 60  $\mu$ g of HA-TC10 and 500  $\mu$ g of LacZ control, or Flag-tagged CAP or CAP $\Delta$ SH3. Cells were starved for 3 h before collection. Top, HA immunoblot of pull-down samples from different time points. Bottom, HA and Flag immunoblots with 50  $\mu$ g of lysate. **b**, 3T3L1 adipocytes were treated with 100 nM insulin for various times. Cells were lysed at 4 °C in the presence of Triton X-100. Triton-insoluble complexes were prepared<sup>10</sup>, and 20  $\mu$ g of the triton-insoluble material was suspended in Laemmli sample buffer, separated by SDS-PAGE, transferred to nitrocellulose and blotted for C3G, CRKII, caveolin I (Cav1) and flotillin. **c**, 3T3L1 adipocytes were transfected by electroporation with LacZ, CAP or CAP $\Delta$ SH3, and stimulated with 100 nM insulin for 5 min. The transfection efficiency was 50%. Cells were lysed at 4 °C in the presence of TritonX-100. Triton-insoluble complexes were suspended in sample buffer, resolved by SDS-PAGE, transferred to nitrocellulose and immunoblotted with anti-C3G and caveolin I antibodies. A representative blot is shown. Films were scanned and quantified, and the ratio of total C3G to caveolin I was calculated. Two asterisks,  $P < 0.01$ ;  $n = 4$ .

formation, Rac and Cdc42 associate specifically with the Cdc42/Rac interactive binding (CRIB) motif of the p21-activated protein kinase 1, Pak1 (refs 15–17). TC10 also binds the CRIB domain of effector proteins in a GTP-dependent manner<sup>13</sup>. To confirm that the Pak1 CRIB domain specifically interacts with these proteins, we transfected Chinese hamster ovary (CHO) cells with HA-epitope tagged wild-type Rac1, Cdc42 or TC10 cDNAs (Supplementary Information Fig. 3). The extracts were incubated with GDP or GTP- $\gamma$ S, and precipitated with a glutathione S-transferase (GST)–Pak1 fusion protein. The fusion protein quantitatively precipitated the GTP- $\gamma$ S-bound forms of Rac, Cdc42 or TC10, showing that GST–Pak1 specifically recognizes the GTP-bound, activated state of these Rho family GTPases.

To examine whether TC10 can be activated in 3T3L1 adipocytes, we electroporated cells with a cDNA encoding HA–TC10 and treated them with insulin (Fig. 2a). Treating the cells with insulin caused the activation of TC10, which was maximal by about 5 min and declined thereafter. The activation of TC10 by insulin was specific, as insulin had no effect on GTP loading of Cdc42 (Supplementary Information Fig. 4). To determine whether activation of TC10 is unique for insulin, we treated HA–TC10-electroporated cells with platelet-derived growth factor (PDGF). In contrast to insulin, TC10 was not activated by PDGF (Supplementary Information Fig. 5). Together, these data provide strong evidence that TC10 is activated by insulin through a specific and unique signalling pathway.

Several studies have shown that PI(3)K activation is necessary, but not sufficient for insulin-stimulated GLUT4 translocation<sup>18–20</sup>. To determine whether TC10 activation requires PI(3)K, 3T3L1 adipocytes were electroporated with HA–TC10, followed by incubation with increasing doses of wortmannin, a PI(3)K inhibitor. Cells were then stimulated with insulin, and lysates were analysed for Akt phosphorylation and TC10 activation. The stimulation of Akt phosphorylation by insulin was completely blocked by incubation with wortmannin<sup>21,22</sup> (Fig. 2b). In contrast, TC10 activation was unaffected by wortmannin, indicating that TC10 activation does not occur downstream of PI(3)K.

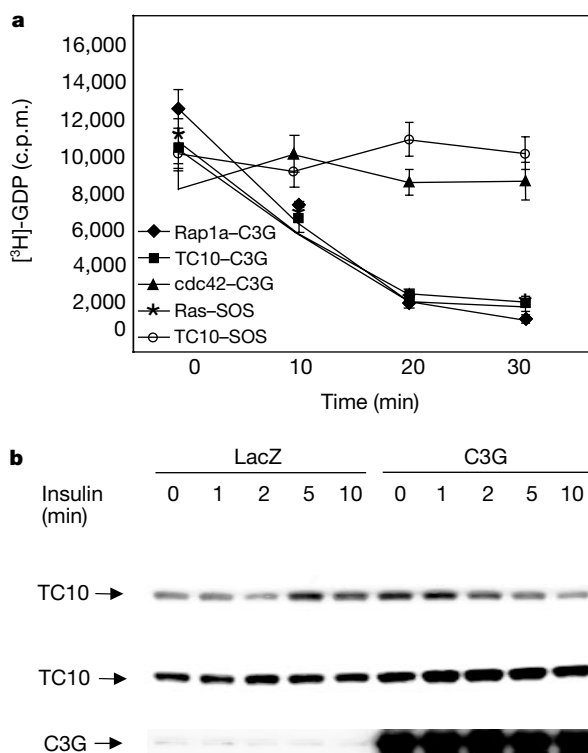
We have reported that the insulin-stimulated tyrosine phosphorylation of Cbl results in its recruitment to flotillin-enriched plasma-membrane subdomains through the adapter protein CAP<sup>10</sup>. Expression of a dominant-interfering CAP mutant (CAP $\Delta$ SH3) has no effect on PI(3)K activation, but prevents the recruitment of tyrosine-phosphorylated Cbl to lipid rafts, and concomitantly inhibits insulin-stimulated glucose uptake and GLUT4 translocation<sup>10</sup>. To assess the potential role of the CAP–Cbl complex in TC10 activation, we co-transfected 3T3L1 adipocytes with HA–TC10 and either wild-type CAP or CAP $\Delta$ SH3 (Fig. 3a). The insulin-stimulated activation of TC10 was unaffected by the co-expression of the wild-type CAP protein; however, expression of CAP $\Delta$ SH3 slightly reduced the basal level of activated TC10, and completely prevented the insulin-stimulated increase in TC10 GTP loading. These data show that the CAP–Cbl axis provides a necessary signal parallel to the PI(3)K pathway for the activation of TC10 by insulin.

Stimulation of the tyrosine phosphorylation of Cbl by insulin generates a specific docking site for the amino-terminal Src homology (SH)-2 domain of the small adapter protein CrkII (ref. 8). Furthermore, the SH3 domain of CrkII interacts directly with the proline-rich domain of the guanyl nucleotide exchange factor C3G<sup>23</sup>. As tyrosine-phosphorylated Cbl is translocated into lipid rafts in response to insulin, we thought that Cbl might also recruit the CrkII–C3G complex into this plasma-membrane region. To test this possibility, 3T3L1 adipocytes were stimulated with insulin, and the lipid raft plasma-membrane domain was isolated as a Triton X-100 insoluble fraction (Fig. 3b). Immunoblotting with an anti-caveolin antibody showed that caveolin I and flotillin were found in the Triton-insoluble fraction, and that their distribution was

unaffected by insulin. In contrast, both CrkII and C3G migrated in the Triton-soluble fraction, but were recruited into the Triton-insoluble fraction in response to insulin. This effect was maximal at 3 min, and was reversed by 30 min. The time course of CrkII and C3G movement to the Triton-insoluble fraction paralleled the insulin-stimulated recruitment of Cbl<sup>10</sup> and the activation of TC10 (Fig. 2). As a control, insulin had no effect on the subcellular localization of the Ras nucleotide exchange factor SOS. This protein was found largely in the Triton-soluble fraction, but small amounts could be seen in the insoluble fraction on overexposure of the gels. Furthermore, the translocation of neither CrkII nor C3G was affected by pretreatment of cells with wortmannin (data not shown).

As expression of the dominant-interfering mutant CAP $\Delta$ SH3 prevented the recruitment of tyrosine-phosphorylated Cbl to the caveolin/flotillin-enriched subdomain<sup>10</sup>, we evaluated its effect on the recruitment of C3G to the Triton-insoluble complex (Fig. 3c). Insulin stimulated a marked recruitment of C3G into the Triton-insoluble fraction in both LacZ- and CAP-expressing cells; however, expression of CAP $\Delta$ SH3 decreased the insulin-stimulated recruitment of C3G by roughly 40%. Taking into account a 50% transfection efficiency, this corresponds to a near complete block of C3G recruitment in the transfected cell population. Thus, insulin stimulates the translocation of C3G into caveolin-enriched compartments through a process that requires CAP function.

The temporal and spatial correlation of TC10 activation with that



**Figure 4** C3G catalyses guanine nucleotide exchange for TC10. **a**, 293T cells were transfected with constructs encoding Myc–C3G and Flag–SOS. Cells were immunoprecipitated with anti-Flag or anti-Myc antibodies, and immunoprecipitates were assayed for exchange activity using GST fusion proteins of [<sup>3</sup>H]GDP-loaded TC10, cdc42, Rap1a and Ras in the presence of GTP. Aliquots were withdrawn at various time points and scintillation counting was performed. Results are the mean  $\pm$  s.e. of triplicate determinations and are representative of three experiments. **b**, 3T3L1 adipocytes were co-electroporated with HA–TC10 and 200  $\mu$ g of a LacZ control or a C3G expression vector. Top, HA immunoblot of pull-down samples from different time points. Bottom, HA and C3G immunoblots with 50  $\mu$ g of lysate.

of C3G recruitment implicates C3G as a possible insulin-regulated guanyl nucleotide exchange factor (GEF) for TC10. To assess the relative GEF activity of C3G for TC10, both C3G and SOS were expressed in 293T cells, and immunoprecipitation was carried out. GTP-loading assays were performed on the immunoprecipitates using Ras, Rap1 or TC10 as substrate (Fig. 4a). As previously reported<sup>24</sup>, SOS had a relatively high activity for Ras, but was a poor GEF for Rap1. In contrast, C3G effectively catalysed the exchange of GTP for GDP on Rap1, but was ineffective with Ras. SOS was also a very weak GEF for TC10, whereas C3G displayed profound GTP/GDP exchange activity for TC10. However, C3G did not show exchange activity with Cdc42 as substrate.

To examine the specific exchange activity of C3G for TC10, we deleted the putative GEF catalytic domain that exhibits sequence similarity to the *Saccharomyces cerevisiae* CDC25 exchange factor (C3G $\Delta$ cdc25). The C3G $\Delta$ cdc25 mutant failed to exchange GTP for GDP on Rap1 as well as TC10 (Supplementary Information Fig. 6). Thus, C3G can function as a specific GEF for TC10 *in vitro*, with activity comparable to that seen with its established target Rap1.

To determine whether C3G can function as a TC10 GEF *in vivo*, we co-transfected 3T3L1 adipocytes with HA–TC10 either with or without an expression vector encoding C3G (Fig. 4b). Transfection with C3G cDNA resulted in a marked increase in C3G protein over endogenous levels, but did not significantly affect the co-expression of HA–TC10. In the absence of C3G overexpression, TC10 was predominantly in the inactive GDP-bound state, but was activated

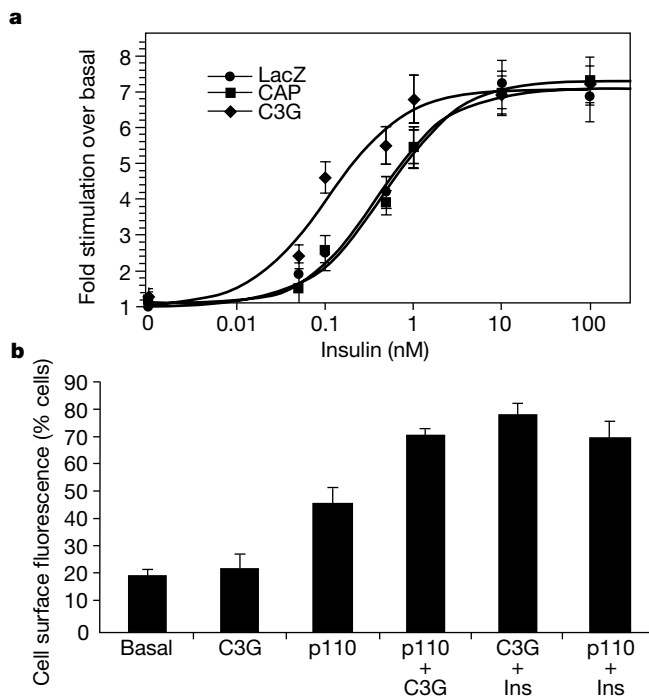
in response to insulin. In cells overexpressing C3G, however, TC10 was quantitatively converted to the active GTP-bound state in the absence of insulin. Under these conditions, TC10 was activated maximally, as insulin did not further increase the amount of GTP-bound TC10.

The activation of TC10 by wild-type C3G suggests that overexpression of C3G may result in its being targeted to lipid rafts in an insulin-independent manner. Consistent with this hypothesis, a significant fraction of the wild-type C3G was detected in the Triton-insoluble fraction, although most of the overexpressed C3G protein was cytosolic (Supplementary Information Fig. 7). Thus, the excess expression of wild-type C3G results in its hormone-independent mistargeting to lipid rafts, where it can catalyse nucleotide exchange on TC10.

Although our data suggest that TC10 activation provides a critical component in the stimulation of glucose transport by insulin, it is also likely that the pathway will not be sufficient for insulin action, as glucose transport also requires the activation of PI(3)K and downstream protein kinases. To examine this issue, we evaluated the impact of TC10 activation on glucose transport by overexpressing C3G in 3T3L1 adipocytes. We transfected cells with C3G as above, and assessed the dose dependence of insulin-stimulated glucose transport. Insulin stimulated glucose uptake in 3T3L1 adipocytes, with an EC<sub>50</sub> (effector concentration at half-maximal response) of about 0.2 nM (Fig. 5a). Expression of C3G resulted in a 4–5-fold leftward shift of the dose–response curve, without any change in the maximal response. This increase in insulin sensitivity suggests that the TC10 pathway has an important role in insulin signal transduction.

Previous studies have shown that expression of a constitutively active PI(3)K catalytic subunit only results in a partial translocation of GLUT4 (refs 18–20). As TC10 activation occurs independently of PI(3)K, we examined whether these two pathways together are sufficient to stimulate GLUT4 translocation fully (Fig. 5b). As typically observed, in the basal state roughly 20% of GLUT4–EGFP-expressing cells displayed a plasma-membrane fluorescence, which increased to 80% after insulin stimulation. Consistent with previous reports<sup>18,19</sup>, co-expression of GLUT4–EGFP with a constitutively active p110 (p110CAAX) resulted in partial GLUT4 translocation that was further stimulated on the addition of insulin. Overexpressing C3G alone did not affect translocation in the basal state, in agreement with glucose transport data (see Fig. 5a); however, overexpressing both C3G and active p110 resulted in 75% of the transfected cells displaying the GLUT4 translocation, which is essentially identical to that induced by insulin. The ability of C3G and p110CAAX to mimic in full the effect of insulin strongly suggests that the combination of these two pathways is sufficient to mediate insulin-stimulated GLUT4 translocation.

These observations lead us to propose a new model for the spatial compartmentalization of signals that regulate glucose transport in adipocytes (Supplementary Information Fig. 8). The insulin receptor can catalyse the tyrosine phosphorylation of a number of substrates, including Cbl and the insulin receptor substrate (IRS) family. On tyrosine phosphorylation, IRS proteins can interact with PI(3)K, leading to the localized generation of phosphatidylinositol triphosphate (InsP<sub>3</sub>) and the subsequent activation/localization of InsP<sub>3</sub>-dependent protein kinases. In parallel, a second pathway is activated by the insulin receptor through the tyrosine phosphorylation of Cbl. The tyrosine-phosphorylated Cbl protein, complexed with the adapter protein CAP, translocates to a lipid raft subdomain in the plasma membrane; the C-terminal domain of CAP binds Cbl, while the N-terminal domain specifically interacts with the lipid raft-localized protein, flotillin. The tyrosine phosphorylation of Cbl directs its association with the CrkII–C3G complex, resulting in its co-recruitment to the lipid raft. Because TC10 is also localized in these plasma-membrane subdomains, C3G is in close enough



**Figure 5** Overexpression of C3G and constitutively active p110 in 3T3L1 adipocytes mimics the effect of insulin on GLUT4 translocation. **a**, 3T3-L1 adipocytes were transfected by electroporation with LacZ, CAP or C3G containing constructs. [<sup>14</sup>C]2-deoxyglucose uptake was measured after exposure to various doses of insulin. Fold stimulation by insulin is calculated as the insulin-stimulated level divided by the basal value. Each point was performed in triplicate; means  $\pm$  s.e. are shown. **b**, 3T3-L1 adipocytes were transfected with 25  $\mu$ g of GLUT4–EGFP, either alone or with 200  $\mu$ g C3G or 200  $\mu$ g p110CAAX (a gift from A. Klippel), or with a combination of both p110CAAX and C3G. Cells were incubated in the absence or presence of 100 nM insulin (Ins) for 30 min. The number of cells displaying plasma-membrane rim fluorescence was then determined. Values are mean  $\pm$  s.e. from three independent experiments in which 75–100 individual cells were counted.



proximity to catalyse the exchange of GTP for GDP, resulting in the activation of TC10.

How does TC10 affect glucose transport? Our data lead us to speculate that TC10 may mediate the activation of two separate downstream pathways by insulin. Insulin-activated TC10 might directly produce a cytoskeletal rearrangement to facilitate the exocytosis of GLUT4. Evidence suggests that actin filaments may have a crucial role in the exocytotic recruitment of GLUT4 to the plasma membrane from an intracellular pool in isolated adipocytes<sup>25</sup>. Alternatively, TC10 might be crucial in the regulation of GLUT4 docking and fusion with the plasma membrane. Insulin can modulate the binding of the vesicle-SNARE protein VAMP2 with the target-SNARE syntaxin 4 to control the docking and fusion of GLUT4 vesicles through a PI(3)K-independent pathway. Thus, it is tempting to speculate that the activation of TC10 is directly upstream of this event. Future studies will be needed to address these possibilities. □

## Methods

### Transfection of CHO/IR and 3T3L1 adipocytes

CHO/IR cells and differentiated 3T3L1 adipocytes were transiently transfected as described<sup>1,10</sup>.

### Guanine nucleotide exchange assay

We transfected 293T cells with Myc-C3G or Flag-SOS plasmid DNA. Transfected cells were cultured for 48 h, gathered and lysed in 25 mM HEPES pH 7.4, 150 mM NaCl, 1.5 mM MgCl<sub>2</sub>, 0.5% Triton X-100 and protease inhibitors. Myc-C3G and Flag-SOS proteins were immunoprecipitated with Myc antibody or Flag antibodies and protein-A/G-Sepharose. Guanine nucleotide exchange assays were performed essentially as described<sup>26</sup>. We constructed C3G and C3GΔcdc25 by cloning the NcoI-XbaI and NcoI-SmaI fragments, respectively, of the coding sequence of C3G into a pCS2-mt vector with a 5' Myc epitope tag.

### Affinity precipitation of TC10 using GST-Pak1 PBD

We modified the method from ref. 15. Cells were incubated with binding buffer (25 mM Tris-HCl, pH 7.5, 1 mM dithiothreitol, 30 mM MgCl<sub>2</sub>, 40 mM NaCl and 0.5% Nonidet P-40) in the presence of 7 μg of GST-Pak1 pzi-binding domain (PBD) agarose (Upstate) for 1 h at 4 °C. The beads were washed three times with 1% Nonidet P-40 washing buffer. The beads were suspended in sample buffer, separated by 4–20% SDS-PAGE, transferred to a nitrocellulose membrane, and blotted with anti-HA monoclonal antibody.

Received 11 September 2000; accepted 16 February 2001.

- Min, J. *et al.* Synip: a novel insulin-regulated syntaxin 4-binding protein mediating GLUT4 translocation in adipocytes. *Mol. Cell* **3**, 751–760 (1999).
- Olson, A. L., Knight, J. B. & Pessin, J. E. Syntaxin 4, VAMP2, and/or VAMP3/ cellubrevin are functional target membrane and vesicle SNAP receptors for insulin-stimulated GLUT4 translocation in adipocytes. *Mol. Cell. Biol.* **17**, 2425–2435 (1997).
- Hausdorff, S. F., Bennett, A. M., Neel, B. G. & Birnbaum, M. J. Different signaling roles of SHPTP2 in insulin-induced GLUT1 expression and GLUT4 translocation. *J. Biol. Chem.* **270**, 12965–12968 (1995).
- Elmendorf, J. S., Chen, D. & Pessin, J. E. Guanosine 5'-O-(3-thiotriphosphate) stimulation of GLUT4 translocation is tyrosine kinase-dependent. *J. Biol. Chem.* **273**, 13289–13296 (1998).
- Holman, G. D. *et al.* Cell surface labeling of glucose transporter isoform GLUT4 by bis-mannose phospholabel. Correlation with stimulation of glucose transport in rat adipose cells by insulin and phorbol ester. *J. Biol. Chem.* **265**, 18172–18179 (1990).
- Piper, R. C., Hess, L. J. & James, D. E. Differential sorting of two glucose transporters expressed in insulin-sensitive cells. *Am. J. Physiol.* **260**, C570–580 (1991).
- Pessin, J. E. & Saltiel, A. R. Signaling pathways in insulin action: molecular targets of insulin resistance. *J. Clin. Invest.* **106**, 165–169 (2000).
- Ribon, V. & Saltiel, A. R. Insulin stimulates tyrosine phosphorylation of the proto-oncogene product of c-Cbl in 3T3-L1 adipocytes. *Biochem. J.* **324**, 839–845 (1997).
- Ribon, V., Printen, J. A., Hoffman, N. G., Kay, B. K. & Saltiel, A. R. A novel, multifunctional c-Cbl binding protein in insulin receptor signaling in 3T3-L1 adipocytes. *Mol. Cell. Biol.* **18**, 872–879 (1998).
- Baumann, C. A. *et al.* CAP defines a second signaling pathway required for insulin-stimulated glucose transport. *Nature* **407**, 202–207 (2000).
- Aspenstrom, P. Effectors for the rho GTPases. *Curr. Opin. Cell Biol.* **11**, 95–102 (1999).
- Kjoller, L. & Hall, A. Signaling to rho GTPases. *Exp. Cell. Res.* **253**, 166–179 (1999).
- Neudauer, C. L., Joberty, G., Tatsis, N. & Macara, I. G. Distinct cellular effects and interactions of the rho-family GTPase TC10. *Curr. Biol.* **8**, 1151–1160 (1998).
- Jou, T. S., Schneeberger, E. E. & Nelson, W. J. Structural and functional regulation of tight junctions by rhoA and rac1 small GTPases. *J. Cell Biol.* **142**, 101–115 (1998).
- Benard, V., Bohl, B. P. & Bokoch, G. M. Characterization of rac and cdc42 activation in chemo-attractant-stimulated human neutrophils using a novel assay for active GTPases. *J. Biol. Chem.* **274**, 13198–13204 (1999).
- Owen, D., Mott, H. R., Laue, E. D. & Lowe, P. N. Residues in cdc42 that specify binding to individual CRIB effector proteins. *Biochemistry* **39**, 1243–1250 (2000).

- Thompson, G., Owen, D., Chalk, P. A. & Lowe, P. N. Delineation of the cdc42/rac-binding domain of p21-activated kinase. *Biochemistry* **37**, 7885–7891 (1998).
- Wiese, R. J., Mastick, C. C., Lazar, D. F. & Saltiel, A. R. Activation of mitogen-activated protein kinase and PI 3-kinase is not sufficient for the hormonal stimulation of glucose uptake, lipogenesis, or glycogen synthesis in 3T3-L1 adipocytes. *J. Biol. Chem.* **270**, 3442–3446 (1995).
- Isakoff, S. J. *et al.* The inability of PI 3-kinase activation to stimulate GLUT4 translocation indicates additional signaling pathways are required for insulin-stimulated glucose uptake. *Proc. Natl Acad. Sci. USA* **92**, 10247–10251 (1995).
- Guilherme, A. & Czech, M. P. Stimulation of IRS-1-associated PI 3-kinase and Akt/ protein kinase B but not glucose transport by beta 1 integrin signaling in rat adipocytes. *J. Biol. Chem.* **273**, 33119–33122 (1998).
- Okada, T., Kawano, Y., Sakakibara, T., Hazeki, O. & Ui, M. Essential role of PI 3-kinase in insulin-induced glucose transport and antilipolysis in rat adipocytes. Studies with a selective inhibitor wortmannin. *J. Biol. Chem.* **269**, 3568–3573 (1994).
- Madge, L. A. & Pober, J. S. A PI-3 kinase/Akt pathway, activated by TNF or IL-1 inhibits apoptosis but not activate NF kappa B in human endothelial cells. *J. Biol. Chem.* **275**, 15458–15465 (2000).
- Knudsen, B. S., Feller, S. M. & Hanafusa, H. Four proline-rich sequences of the guanine-nucleotide exchange factor C3G bind with unique specificity to the first Src homology 3 domain of crk. *J. Biol. Chem.* **269**, 32781–32787 (1994).
- Gotoh, T. *et al.* Identification of Rap1 as a target for the crk SH3 domain binding guanine nucleotide-releasing factor C3G. *Mol. Cell. Biol.* **15**, 6746–6753 (1995).
- Omata, W., Shibata, H., Kuniaki, L. L. & Kojima, I. Actin filaments play a critical role in insulin-induced exocytotic recruitment but not in endocytosis of GLUT4 in isolated rat adipocytes. *Biochem. J.* **346**, 321–328 (2000).
- Hori, Y., Beeler, J. F., Sakaguchi, M., Tachibana, M. & Miki, T. A novel oncogene, *ost*, encodes a guanine nucleotide exchange factor that potentially links Rho and Rac signaling pathways. *EMBO J.* **13**, 4776–4786 (1994).
- Thurmond, D. C., Kanzaki, M., Khan, A. H. & Pessin, J. E. Munc18c function is required for insulin-stimulated plasma membrane fusion of GLUT4 and insulin-responsive amino peptidase storage vesicles. *Mol. Cell. Biol.* **20**, 379–388 (2000).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Correspondence and requests for materials should be addressed to A.R.S. (e-mail: saltiel@umich.edu).

## The neuronal repellent Slit inhibits leukocyte chemotaxis induced by chemotactic factors

Jane Y. Wu\*, Lili Feng†, Hwan-Tae Park\*, Necat Havlioglu\*, Leng Wen‡, Hao Tang\*, Kevin B. Bacon§, Zhi-hong Jiang\*, Xiao-chun Zhang\* & Yi Rao‡

\* Departments of Pediatrics, and Molecular Biology and Pharmacology, and ‡ Department of Anatomy and Neurobiology, Washington University School of Medicine, Box 8108, 660 S. Euclid Avenue, St Louis, Missouri 63110, USA  
† Department of Medicine, Division of Nephrology, Baylor College of Medicine, One Baylor Plaza, N730 Houston, Texas 77030, USA  
§ Department of Biology I, Bayer Yakuhin Ltd, 6-5-1-3 Kunimidai, Kizu-cho, Soraku-gun, Kyoto, Japan

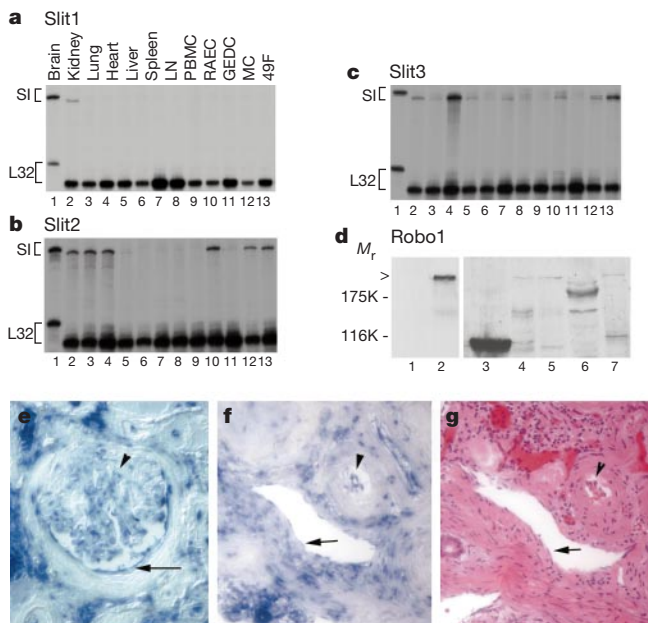
Migration is a basic feature of many cell types in a wide range of species<sup>1</sup>. Since the 1800s, cell migration has been proposed to occur in the nervous and immune systems<sup>2,3</sup>, and distinct molecular cues for mammalian neurons and leukocytes have been identified. Here we report that Slit, a secreted protein previously known for its role of repulsion in axon guidance and neuronal migration, can also inhibit leukocyte chemotaxis induced by chemotactic factors. Slit inhibition of the chemokine-induced chemotaxis can be reconstituted by the co-expression of a chemokine receptor containing seven transmembrane domains and Roundabout (Robo), a Slit receptor containing a single transmembrane domain. Thus, there is a functional interaction between single and seven transmembrane receptors. Our results reveal the activity of a neuronal guidance cue in regulating leukocyte migration and indicate that there may be a general

**conservation of guidance mechanisms underlying metazoan cell migration. In addition, we have uncovered an inhibitor of leukocyte chemotaxis, and propose a new therapeutic approach to treat diseases involving leukocyte migration and chemotactic factors.**

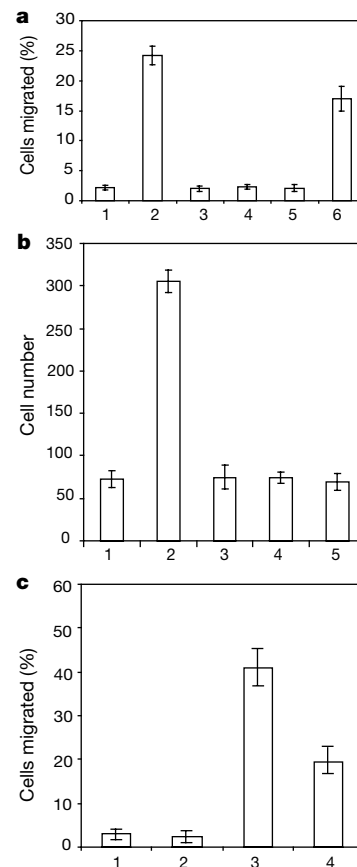
Leukocyte chemotaxis is one of the best-characterized models of cell migration in adult mammals<sup>4–12</sup>. Work in the past 20 years has demonstrated the importance of the chemokine family in leukocyte chemotaxis<sup>4–9</sup>. These structurally related small proteins regulate leukocyte migration and function<sup>4–9</sup>. Neuronal migration is an important step in neural development in the nervous system<sup>10–12</sup>; however, it is not clear whether mechanisms operating in the nervous system are conserved in the immune system. Leukocytes migrate at a rate much faster than neurons: the behaviour of leukocytes can be monitored in seconds, whereas that of the neurons is monitored in tens of minutes and hours. The cellular environment in which leukocytes migrate also seems to be different from that for axons and neurons. The morphology of migrating neurons, which have a rather long leading process, differs from that of leukocytes. At the molecular level, all neuronal guidance cues

bind to receptors containing a single transmembrane domain. By contrast, G-protein-coupled receptors (GPCRs) containing seven transmembrane domains are thought to be the sole receptors mediating responses not only to all chemokines, but also to other chemotactic factors either for leukocytes<sup>5</sup> or for the social amoebae *Dictyostelium*<sup>1</sup>. It is therefore not obvious whether neuronal guidance cues can have roles in the haematopoietic or immune systems.

Studies on the *slit* gene family<sup>13,14</sup> have shown that secreted Slit proteins can guide both axon projection<sup>15–18</sup> and neuronal migration<sup>19–22</sup>. We previously observed embryonic expression of Slit and Robo outside the mammalian nervous system<sup>17,22</sup>, which led us to speculate potential roles for Slit in several systems including the immune system<sup>19,20</sup>. Previous studies by us and others have focused on embryonic expression of *slit* and *robo*. We have now used RNase protection assays to examine the expression of three known *slit* genes in adult tissues and cell lines (Fig. 1a–c). Each *slit* gene has a distinct expression pattern: *slit1* is specific to the brain, but *slit2* and *slit3* are expressed in the brain as well as other tissues. The expression level of *slit2* in the kidney and the lung is comparable to that in the brain. Expression of *slit3* is the highest in the lung and



**Figure 1** Expression of Slit and Robo in adult tissues. **a–c**, RNase protection assays (RPAs) were used to determine the expression of three *slit* genes in adult rats. Total RNA was prepared from different rat tissues and cell lines, as indicated (lanes 2–13). Total RNA (5  $\mu$ g) was used in RPAs with probes specific for *slit1*, *slit2* and *slit3* genes (SI). The rat L32 gene probe controlled for RNA input (lower band in each panel). Probes in lane 1 contain polylinker regions and are longer than the protected bands. LN, lymph node; PBMC, peripheral blood mononuclear cells; RAEC, rat endothelial cell line; GEDC, glomerular endothelial cells from the rat kidney; MC, rat kidney mesangial cells; 49F, rat kidney fibroblasts. **d**, Expression of Robo1 protein detected with anti-Robo1 antibodies. Arrowhead indicates full-length Robo1. Lane 1, HEK cells; lane 2, Robo1-transfected HEK cells; lane 3, a strong band with lower relative molecular mass ( $M_r$ ) than full-length Robo1 was detected reproducibly in rat thymus, indicating a crossreacting band or a proteolytically cleaved product; lane 4, HL-60 cells; lane 5, neutrophils differentiated from HL-60 cells; lane 6, rat PBMCs: no full-length Robo1 was detected but several lower  $M_r$  bands were visible; lane 7, rat lymph node. **e**, *slit2* messenger RNA distribution detected by *in situ* hybridization in the glomerulus of adult human kidney. Arrowhead indicates mesangial cells; arrow indicates epithelial cells in Bowman's capsule. In the glomerulus, most of the positive cells are epithelial cells, but some endothelial cells can be seen. **f**, *slit2* mRNA expression in vascular endothelial cells in the human kidney. Arrow indicates endothelial cells in a venule; arrowhead indicates endothelial cells in an arteriole. **g**, Haematoxylin–eosin staining of a section of the same kidney as that shown in **f**. These two sections are not immediately next to each other, but corresponding regions are conveniently identified by landmarks in the sections.



**Figure 2** Effect of Slit on leukocyte chemotaxis induced by SDF-1 $\alpha$ . **a**, Transwell migration of rat lymphocytes. Control or SDF-containing media were added to the lower chamber. Cells that migrated to the lower chamber were counted, and are expressed as a percentage of the cells added to the upper chamber. 1, control; 2, 10 nM SDF-1 $\alpha$ ; 3, 100 pM SLIT2; 4, 10 pM SLIT2; 5, 10 nM of SDF-1 $\alpha$  and 100 pM of SLIT2; 6, 10 nM of SDF-1 $\alpha$  and 10 pM of SLIT2. **b**, Rat lymphocytes were examined in transfilter assays in the presence of SDF-1 $\alpha$  (10 nM). SLIT2 (100 pM) was added to the lower chamber, the upper chamber or both upper and lower chambers. 1, control; 2, SDF-1 $\alpha$ ; 3, both SDF-1 $\alpha$  and SLIT2 in the lower chamber; 4, SDF-1 $\alpha$  in the lower chamber and SLIT2 in the upper chamber; 5, SDF-1 $\alpha$  in the lower chamber and SLIT2 in both chambers. **c**, Slit inhibition of chemotaxis induced by fMLP. HL-60 cells differentiated into neutrophil-like cells after treatment with DMSO. Chemotaxis was observed in transwell assays. 1, control; 2, SLIT2; 3, fMLP; 4, SLIT2 and fMLP.

also detectable in the kidney, the brain, the heart, the spleen and the lymph nodes. When a number of cell lines were examined, *slit2* and *slit3* were found in the rat endothelial cells, mesangial cells and fibroblasts from the rat kidney. *In situ* hybridization of human kidneys revealed the expression of *slit2* in mesangial and epithelial cells in the glomeruli; in epithelial cells in the tubules (Fig. 1e); and in endothelial cells of both arterioles and venules in the kidney (Fig. 1f).

Anti-Robo1 antibodies detected Robo1 in the lymph nodes, the thymus and neutrophils differentiated from HL-60 cells (Fig. 1d). By polymerase chain reaction with reverse transcription (RT-PCR), *robo1* was detected in the thymus, the spleen, the lymph nodes, the liver, the kidney and the heart (data not shown), whereas *robo2* was found in the spleen, the thymus, the liver, the lung and the kidney (data not shown).

To investigate whether Slit could affect chemotaxis, we used standard transwell and transfilter assays in which a chemokine was placed in the lower chamber and the leukocytes were placed in the upper chamber. We then analysed the migration of leukocytes into the lower chamber or onto the underside of the filter. Conditioned serum-free media from human embryonic kidney (HEK) cells expressing human and *Xenopus* Slit2 (SLIT2 and *Xenopus* Slit2) proteins<sup>17</sup>, as well as SLIT2 and *Xenopus* Slit2 proteins purified from serum-containing conditioned media, were tested.

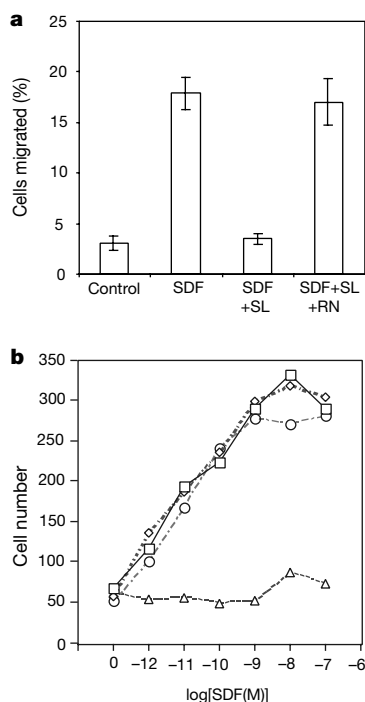
As expected, the chemokine stromal-derived factor (SDF)-1 induced leukocyte chemotaxis (Fig. 2a, b). Lymphocytes isolated from the lymph nodes were chemotactic towards 10 nM of SDF-1

(Fig. 2a), but chemotaxis induced by SDF-1 was reduced by the presence of 100 pM of SLIT2 in the lower chamber (Fig. 2a). This effect was dose dependent; 10 pM of SLIT2 was not effective in reducing chemotaxis induced by SDF-1 (Fig. 2a). The effect of SLIT2 in the lower chamber might be due to either repulsion or inhibition. To test whether Slit could inhibit leukocyte chemotaxis, we added SLIT2 to either the upper or the lower chamber, and SDF-1 to the lower chamber (Fig. 2b). Slit reduced SDF-1-induced chemotaxis when it was present in the upper chamber, and even when it was present in both the upper and the lower chambers. These results indicate that Slit inhibits chemotaxis induced by SDF-1.

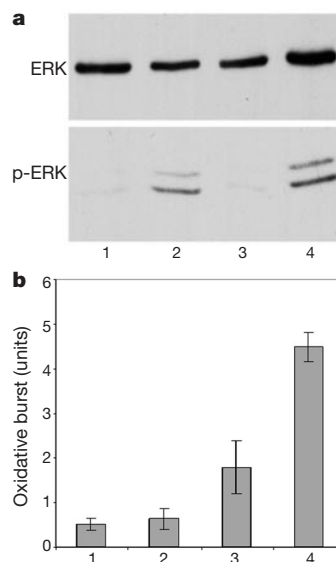
To test whether Slit could affect chemotaxis induced by other chemotactic factors, we used the bacterial product *N*-formyl peptide f-Met-Leu-Phe (fMLP). fMLP attracts neutrophil-like cells differentiated from HL-60 cells<sup>23</sup>. We found that SLIT2 could inhibit the migration of differentiated HL-60 cells induced by fMLP (Fig. 2c).

Robo is a protein with a single transmembrane domain<sup>24,25</sup>, serving as a Slit receptor in axon guidance<sup>15–17</sup> and neuronal migration<sup>19,21</sup>. To investigate whether Robo mediates leukocyte responses to Slit, we used RoboN, a fragment of Robo that contains only the extracellular part of the Robo protein<sup>19</sup>. The addition of RoboN abolished the inhibitory effect of Slit on leukocyte chemotaxis (Fig. 3a). Together with the expression of Robo in the haematopoietic system, these results provide evidence for the involvement of Robo in mediating Slit responses in leukocytes, supporting a conserved guidance mechanism for cell migration.

The receptor for SDF-1 is CXCR4, a GPCR with seven transmembrane domains. To determine whether Robo and CXCR4 are sufficient to mediate the functional interaction between Slit and SDF-1, we introduced Robo and CXCR4 separately or together into HEK cells (Fig. 3b). When transfected with CXCR4, HEK cells migrated towards SDF-1. However, Slit significantly reduced SDF-1-induced chemotaxis of HEK cells expressing both *robo1* and CXCR4, showing that the functional interaction between Slit and



**Figure 3** Robo is involved in mediating Slit inhibition of chemotaxis. **a**, The effect of Slit is inhibited by RoboN. Chemotaxis was analysed by using rat lymph node cells in transwell assays. SLIT2 (SL) and RoboN (RN) were added to the upper well and SDF-1 $\alpha$  was added to the lower well. **b**, Expression of CXCR4 and Robo in HEK cells can reconstitute Slit inhibition of SDF-1 $\alpha$ -induced chemotaxis. HEK migration was measured with the transfilter assay in microchemotaxis chambers using HEK cells expressing CXCR4, or both CXCR4 and rat Robo1 in the presence of different concentrations of SDF-1 $\alpha$ . Control vehicle or 100 pM of purified Slit protein was added. Squares, migration of HEK cells expressing CXCR4 in response to SDF-1; diamonds, migration of HEK cells expressing CXCR4 in response to SDF-1 $\alpha$  in the presence of Slit; circles, migration of HEK cells expressing Robo and CXCR4 in response to SDF-1 $\alpha$ ; triangles, migration of HEK cells expressing Robo and CXCR4 in response to SDF-1 $\alpha$  in the presence of Slit.



**Figure 4** Lack of general inhibition by Slit. **a**, Neutrophils differentiated from HL-60 were treated with fMLP, or fMLP and SLIT2. Phosphorylation of ERK was examined by an anti-phospho-ERK antibody. Top, anti-ERK staining; bottom, anti-phospho-ERK staining. fMLP could induce ERK phosphorylation, which was not inhibited by the addition of SLIT2. Lane 1, control; lane 2, fMLP; lane 3, SLIT2; lane 4, fMLP and SLIT2. **b**, Oxidative burst was shown by SOD-inhibitable production of superoxide, and 1 unit was defined as nmol per  $2 \times 10^6$  cells per 15 min. 1, control; 2, SLIT2; 3, fMLP; 4, SLIT 2 and fMLP. Results shown are representative of three experiments, each with triplicate samples.



the chemokine SDF-1 could be reconstituted in HEK cells by the expression of both Robo and CXCR4.

We next considered whether the inhibitory effect of Slit on migratory responses was due to a general inhibitory or toxic effect on leukocytes. We tested whether Slit could inhibit two activities of fMLP that are not related to cell migration. fMLP activates the extracellular signal-regulated kinases 1 and 2 (ERK1 and ERK2) by causing their phosphorylation (Fig. 4a)<sup>26,27</sup>; however, Slit did not inhibit fMLP-induced ERK phosphorylation (Fig. 4a). fMLP also causes the oxidative burst of neutrophils<sup>28</sup> (Fig. 4b); Slit did not cause oxidative burst, nor did it inhibit the oxidative burst induced by fMLP (Fig. 4b). In fact, Slit increased the oxidative burst induced by fMLP, which is consistent with the previous observation that regulation of the actin cytoskeleton can prime cells for oxidative burst. These results indicate that Slit inhibits chemotaxis, but not all other functions of leukocytes.

Our findings that Slit can inhibit chemotaxis of leukocytes induced by chemotactic factors have implications in the conservation of molecular mechanisms of cell migration between species, and in biological and therapeutic regulations of immune responses and other physiological or pathological processes involving chemotactic factors. We suggest that neuronal guidance cues such as netrins, semaphorins and ephrins may also guide leukocyte migration and regulate biological processes involving chemotactic factors. Other molecules involved in neuronal positioning, such as Reelin<sup>12</sup>, should also be tested for their potential roles in leukocyte migration. Our results in HEK cells indicate a functional interaction between the single transmembrane receptor Robo and the GPCRs, suggesting a mechanism for regulating signal transduction pathways mediated by GPCRs and heterotrimeric G proteins.

Past research in leukocyte chemotaxis has been focused mainly on chemoattractants and positive regulation. Although there were reports of virally produced inhibitors of chemokines<sup>29,30</sup>, there is no known endogenous negative factors. Our results reveal that there are negative regulators of leukocyte chemotaxis available endogenously, which may be important in physiological and pathological situations. This expands the scope and the perspective of research in leukocyte biology. Chemokines have several roles, including inflammatory responses, leukocyte activation, lymphocyte trafficking and lymphoid organ homeostasis, tissue injury, haematopoiesis, atherosclerosis, allogeneic transplant rejection, angiogenesis, virally induced vascular diseases, cardiac morphogenesis, and tumour development. In all of these situations except tumorigenesis, inhibition of chemokine signalling is therapeutically beneficial, and much effort has been directed towards obtaining reagents that can block chemokines or their receptors. Potential applications of Slit should be tested.

Our studies of Slit broaden the search for chemokine inhibitors. Because the inhibition of receptor signalling can block HIV infection, Slit inhibition of chemokine receptors including CXCR4 (and CCR5; data not shown) suggests that Slit cannot be ruled out as a useful reagent in inhibiting HIV infection; thus, the therapeutic potentials for Slit and other negative factors are an attractive area of further research. Our unpublished data indicate that *in vivo* application of Slit protein attenuates crescentic glomerulonephritis in an animal model involving chemokine-induced leukocyte infiltration. In summary, our findings of Slit inhibition of leukocyte chemotaxis induced by chemotactic factors not only shed light on the fundamental conservation of mechanisms guiding cell migration, but also open up new areas for future investigations. □

## Methods

### RNAse protection assay

We used 5 µg total RNA for each sample prepared from tissues or cultured cells in RNAse protection assays. Riboprobes specific for rat *slit1*, *slit2* and *slit3* genes were prepared, and RNAse protection assay was performed using a kit (Torrey Pines Biologicals) according to the manufacturer's protocols with corresponding probes labelled with [<sup>32</sup>P]UTP.

### Isolation of leukocytes and assays with HL-60 cells

Leukocytes were isolated from rat lymph nodes, spleen or peripheral blood according to standard protocols. Cells were kept at 4 °C in media containing 5% fetal calf serum (FCS) until use; and cells were used within 8 h after isolation.

We grew the HL-60 cell line under standard conditions with RPMI-1640 media supplemented with 10% heat-inactivated FCS<sup>23</sup>. Cells induced with 1.2% dimethyl sulphoxide (DMSO) were obtained by seeding HL-60 cells at  $3 \times 10^6$  ml<sup>-1</sup> in growth media and culturing for 4–6 days (ref. 23). ERK Phosphorylation induced by 1 min of fMLP stimulation was examined with anti-phospho ERK antibody<sup>26</sup>. In Fig. 4a, serum-free conditioned medium from control HEK cells was used as control in lane 1; 100 nM of fMLP in the presence of serum-free conditioned medium from control HEK cells was used in lane 2; serum-free conditioned medium from SLIT2-secreting HEK cells was used in lane 3; 100 nM of fMLP in the presence of serum-free conditioned medium from SLIT2 secreting HEK cells was used in lane 4. In Fig. 4b, oxidative burst was quantified according as described<sup>28</sup> by measuring the superoxide dismutase (SOD)-inhibitable reduction of cytochrome c. Briefly, differentiated HL-60 cells ( $2 \times 10^6$ ) containing 75 µM cytochrome c with or without SOD (30 µg) were incubated with fMLP (100 nM) for 15 min at 37 °C. The reaction was stopped by submerging the tubes in ice. The tubes were centrifuged, and the absorbency of the supernatant at 550 nm was measured in a spectrophotometer.

### Assays for leukocyte chemotaxis

Leukocyte chemotaxis was measured by transwell and transfilter assays. Isolated leukocytes were resuspended at a concentration of  $4 \times 10^6$  cells per ml in 50% DMEM, 50% M199, 5% heat-inactivated fetal bovine serum (FBS). Unless otherwise specified, 10 nM chemokines were placed at lower wells of chemotaxis chambers. Chemotaxis results are representative of at least three independent experiments performed in triplicates.

In the transwell assay, cells (100 µl) were put in inserts (5 µm in pore size) (Costar) placed in 24-well dish containing 600 µl of culture media per well with a chemokine or the Slit protein. After incubation at 37 °C for 1.5–3 h, cells migrated through the insert filter into the lower well were collected and counted.

In the transfilter assay, we placed different chemokines and Slit in the lower wells of a 48-well chemotaxis chamber (Neuroprobe, Cabin John, MD) and separated them by a polyvinylpyrrolidone-free polycarbonate filter (5 or 8 µm in pore size). Isolated leukocytes or transfected HEK cells (50 µl) were put in the upper wells. After incubating at 37 °C for 2 h, cells on the top surface of the filter were removed and cells that had migrated through the filter onto the undersurface were fixed in methanol and stained. We expressed the transfilter migration of cells as cell number per five high-power (×400) fields.

### HEK cell culture

We grew HEK cells in DMEM supplemented with 10% FBS. An HEK cell line expressing the extracellular domain of rat Robo1 protein tagged at the carboxy terminus with an haemagglutinin A (HA) tag (RoboN)<sup>19</sup> was cultured in DMEM and 5% FBS. The RoboN-containing media was collected 3–4 d after cells became confluent. The RoboN-containing media was diluted to roughly 1 nM and used in cell migration assay. We also established a stable HEK cell line expressing rat CXCR4.

Received 30 August 2000; accepted 19 February 2001.

- Devreotes, P. N. & Zigmond, S. H. Chemotaxis in eukaryotic cells: a focus on leukocytes and *Dictyostelium*. *Annu. Rev. Cell Biol.* **4**, 649–686 (1988).
- McCutcheon, M. Chemotaxis in leukocytes. *Physiol. Rev.* **26**, 319–336 (1946).
- Bentivoglio, M. & Mazzarello, P. The history of radial glia. *Brain Res. Bull.* **49**, 305–315 (1999).
- Springer, T. A. Traffic signals for lymphocyte recirculation and leukocyte emigration: the multistep paradigm. *Cell* **76**, 301–314 (1994).
- Murphy, P. M. The molecular biology of leukocyte chemoattractant receptors. *Annu. Rev. Immunol.* **12**, 593–633 (1994).
- Baggiolini, M. Chemokines and leukocyte traffic. *Nature* **392**, 565–568 (1998).
- Luster, A. D. Chemokines—chemotactic cytokines that mediate inflammation. *N. Engl. J. Med.* **338**, 436–445 (1998).
- Cyster, J. G. Chemokines and cell migration in secondary lymphoid organs. *Science* **286**, 2098–2102 (1999).
- Campbell, J. J. & Butcher, E. C. Chemokines in tissue-specific and microenvironment-specific lymphocyte homing. *Curr. Opin. Immunol.* **12**, 336–341 (2000).
- Rakic, P. Principles of neural cell migration. *Experientia* **46**, 882–891 (1990).
- Hatten, M. E. & Heintz, N. in *Fundamentals of Neuroscience* (ed. Zigmond, M.) 451–479 (Academic, New York, 1998).
- Rice, D. S. & Curran, T. Mutant mice with scrambled brains: understanding the signaling pathways that control cell positioning in the CNS. *Genes Dev.* **13**, 2758–2773 (1999).
- Nusslein-Volhard, C., Wieschaus, E. & Kluding, H. Mutations affecting the pattern of the larval cuticle in *Drosophila melanogaster*. I. Zygotic loci on the second chromosome. *Roux's Arch. Dev. Biol.* **193**, 267–282 (1984).
- Rothberg, J. M., Hartley, D. A., Walther, Z. & Artavanis-Tsakonas, S. *slit*: an EGF-homologous locus of *D. melanogaster* involved in the development of the embryonic central nervous system. *Cell* **55**, 1047–1059 (1988).
- Kidd, T., Bland, K. S. & Goodman, C. S. Slit is the midline repellent for the Robo receptor in *Drosophila*. *Cell* **96**, 785–794 (1999).
- Brose, K. *et al.* Evolutionary conservation of the repulsive axon guidance function of Slit proteins and of their interactions with Robo receptors. *Cell* **96**, 795–806 (1999).
- Li, H. S. *et al.* Vertebrate Slit, a secreted ligand for the transmembrane protein roundabout, is a repellent for olfactory bulb axons. *Cell* **96**, 807–818 (1999).
- Wang, K. H. *et al.* Purification of an axon elongation- and branch-promoting activity from brain identifies a mammalian Slit protein as a positive regulator of sensory axon growth. *Cell* **96**, 771–784 (1999).

19. Wu, W. *et al.* Directional guidance of neuronal migration in the olfactory system by the secreted protein Slit. *Nature* **400**, 331–336 (1999).
20. Hu, H. Chemorepulsion of neuronal migration by Slit2 in the developing mammalian forebrain. *Neuron* **23**, 703–711 (1999).
21. Zhu, Y., Li, H. S., Zhou, L., Wu, J. Y. & Rao, Y. Cellular and molecular guidance of GABAergic neuronal migration from the striatum to the neocortex. *Neuron* **23**, 473–485 (1999).
22. Yuan, W. *et al.* The mouse Slit family: secreted ligands for Robo expressed in patterns that suggest a role in morphogenesis and axon guidance. *Dev. Biol.* **212**, 290–306 (1999).
23. Collins, S. J., Fuscetti, F. W., Gallagher, R. E. & Gallo, R. C. Normal functional characteristics of cultured human promyelocytic leukemia cells (HL60) after induction of differentiation by dimethylsulfoxide. *J. Exp. Med.* **149**, 969–974 (1979).
24. Kidd, T. *et al.* Roundabout controls axon crossing of the CNS midline and defines a novel subfamily of evolutionarily conserved guidance receptors. *Cell* **92**, 205–215 (1998).
25. Zallen, J. A., Yi, B. A. & Bargmann, C. I. The conserved immunoglobulin superfamily member SAX-3/Robo directs multiple aspects of axon guidance in *C. elegans*. *Cell* **92**, 217–227 (1998).
26. Krump, E., Sanghera, J. S., Pelech, S. L., Furuya, W. & Grinstein, S. Chemotactic peptide N-formyl-Met-Leu-Phe activation of p38 mitogen-activated protein kinase (MAPK) and MAPK-activated protein kinase-2 in human neutrophils. *J. Biol. Chem.* **272**, 937–944 (1997).
27. Downey, G. P. *et al.* Importance of Mek in neutrophil microbicidal responsiveness. *J. Immunol.* **160**, 434–443 (1998).
28. O'Brien, P. J. Superoxide production. *Methods Enzymol.* **105**, 370–379 (1984).
29. Smith, C. A. *et al.* Poxvirus genomes encode a secreted, soluble protein that preferentially inhibits beta chemokine activity yet lacks sequence homology to known chemokine receptors. *Virology* **236**, 316–327 (1997).
30. Kleidal, T. N. *et al.* A broad-spectrum chemokine antagonist encoded by Kaposi's sarcoma-associated herpes virus. *Science* **277**, 1656–1659 (1997).

## Acknowledgements

The amount of data presented and the number papers cited have been limited by space constraints. We are grateful to X. He for help with FACS; to W. Smith for providing us with the HL60 cell line; to the NIH for grant support (to J.Y.W., L.F. and Y.R.); to the John Merck fund, the Klingenstein foundation, and the Leukemia Society of America for scholar awards (to Y.R. and J.Y.W.).

Correspondence and requests for materials should be addressed to J.Y.W. (e-mail: jwu@pcg.wustl.edu), L.F. (e-mail: lfeng@bcm.tmc.edu) or Y.R. (e-mail: raoyi@thalamus.wustl.edu).

## Phototropin-related NPL1 controls chloroplast relocation induced by blue light

Jose A. Jarillo\*, Halina Gabrys†, Juan Capel\*‡, Jose M. Alonso\*§, Joseph R. Ecker\*§ & Anthony R. Cashmore\*

\* Plant Science Institute, Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6018, USA

† Institute of Molecular Biology, Jagiellonian University, 31-120 Krakow, Poland

‡ Departamento de Biología Aplicada, Universidad de Almería, 04120 Almería, Spain

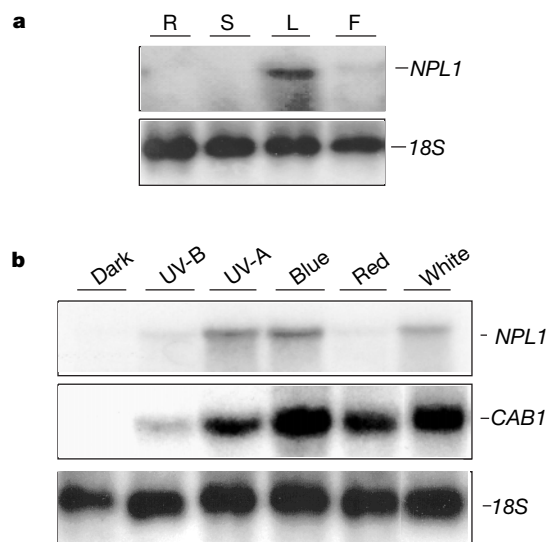
In photosynthetic cells, chloroplasts migrate towards illuminated sites to optimize photosynthesis and move away from excessively illuminated areas to protect the photosynthetic machinery<sup>1</sup>. Although this movement of chloroplasts in response to light has been known for over a century, the photoreceptor mediating this process has not been identified. The *Arabidopsis* gene *NPL1* (ref. 2) is a paralogue of the *NPH1* gene, which encodes phototropin, a photoreceptor for phototropic bending<sup>3</sup>. Here we show that *NPL1* is required for chloroplast relocation induced by blue light. A loss-of-function *npl1* mutant showed no chloroplast avoidance response in strong blue light, whereas the accumulation of chloroplasts in weak light was normal. These results indicate that *NPL1* may function as a photoreceptor mediating chloroplast relocation.

§ Present address: Plant Biology Laboratory, The Salk Institute for Biological Studies, La Jolla, California 92037, USA.

In most plants, chloroplast movement is controlled by blue light and is affected by light direction, as well as wavelength and irradiance<sup>4</sup>. When leaves are exposed to low fluence rates of light, the chloroplasts gather at the periclinal walls, perpendicular to the leaf surface, to maximize photosynthesis. Under high fluence rates of light, chloroplast damage is minimized by movement of the chloroplasts towards the anticlinal walls parallel to the leaf surface. From studies of action spectra, this response appears to involve a flavin<sup>1</sup>, and in some cases phytochrome appears to be involved (see ref. 5 and references therein).

The *NPL1* gene<sup>2</sup> encodes a protein that is similar in sequence to *Arabidopsis* NPH1 (ref. 3), as well as the oat, rice, maize and fern NPH1-related proteins<sup>5–9</sup> (see Supplementary Information Fig. 1). The *Arabidopsis* NPH1 and NPL1 proteins have overall 58% identity and 67% similarity. The amino-terminal LOV1 and LOV2 domains are highly conserved<sup>3</sup>. Furthermore, the cysteine residues, which are important in the photochemical reactions involving flavin mononucleotide (FMW) bound to the LOV domains of NPH1 (ref. 10), are also conserved in NPL1 and the other phototropin family members. This observation indicates that NPL1 (and the other phototropin-like proteins) may also bind FMN as a chromophore. Likewise, high similarity (almost 80%) between phototropin and NPL1 is found in the eleven kinase signature domains that constitute almost the entire carboxy-terminal part of the phototropin-like proteins<sup>11</sup> (see Supplementary Information Fig. 1).

To analyse the expression pattern of the *NPL1* transcript, we extracted total RNA from roots, stems, leaves and flowers of *Arabidopsis* plants. *NPL1* messenger RNA was expressed in all of these organs except roots, with the highest expression detected in leaves (Fig. 1a). Because *NPL1* RNA was most abundant in green tissues, we investigated the effect of light on its expression. The expression of *NPL1* mRNA increased when *Arabidopsis* seedlings were grown under continuous white light, ultraviolet-A or blue light (Fig. 1b). Expression of the rice gene *OsNPH1b*, an apparent orthologue of *NPL1*, is also induced by light, whereas *OsNPH1a*, an orthologue of *NPH1*, is downregulated by light<sup>8</sup>.



**Figure 1** Tissue expression and light inducibility of *NPL1* mRNA. **a**, RNA-blot hybridizations of total RNA (10 µg) extracted from roots (R), stems (S), leaves (L) and flowers (F) of 3–4-week-old plants using a *NPL1* probe. **b**, Total RNA (10 µg) from 6-day-old etiolated seedlings exposed to different light treatments were fractionated by electrophoresis, blotted and hybridized with a *NPL1* probe. RNA was prepared from dark-grown seedlings which were subsequently exposed to ultraviolet-B, ultraviolet-A, blue, red or white light. As a control for the light treatments, the blot was stripped and hybridized with a probe that detects *CAB1* mRNA<sup>19</sup>. As a loading control, blots were reprobed with the 18S ribosomal RNA gene.

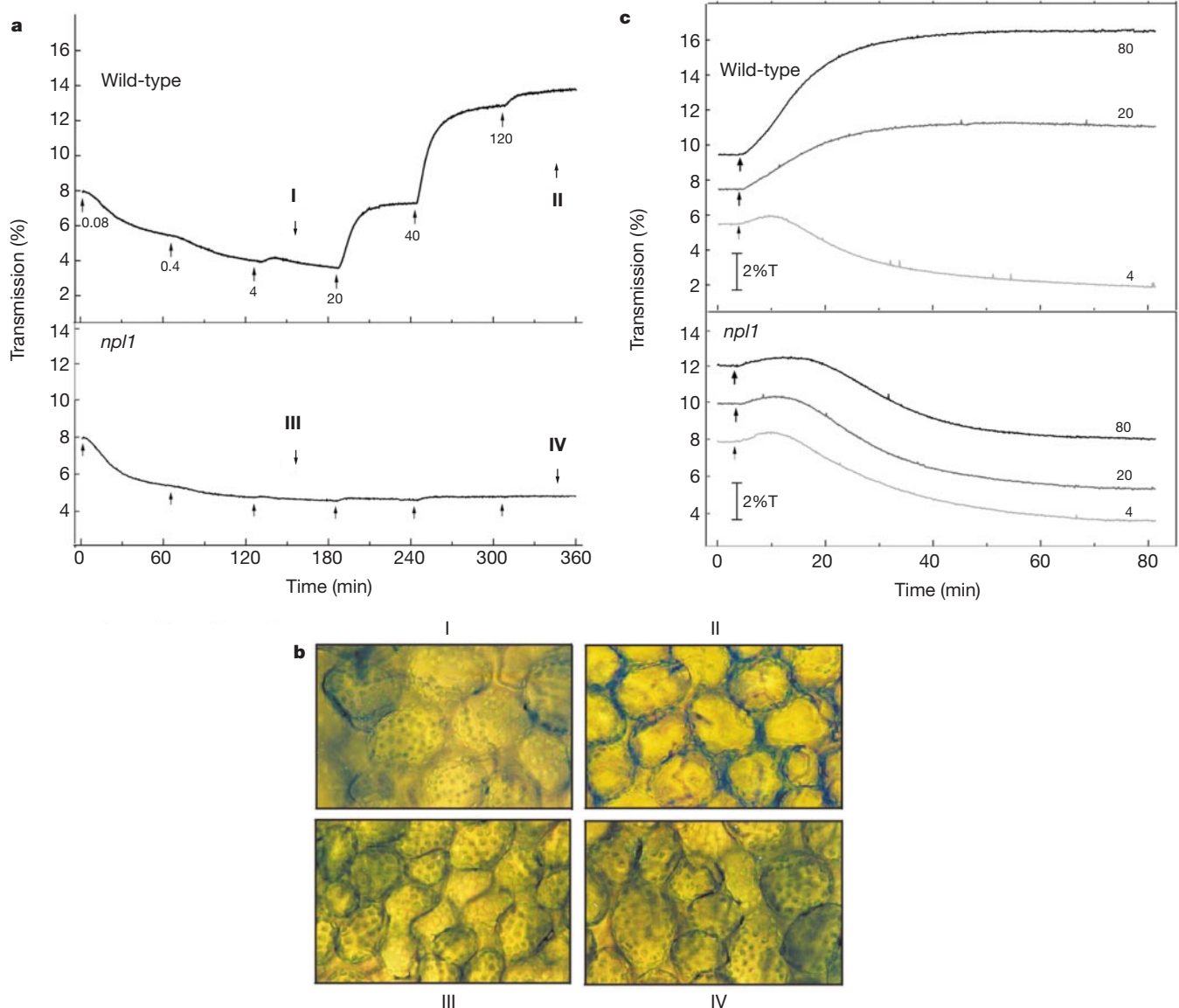
To define the function of *NPL1*, we took a reverse genetic approach. We used polymerase chain reaction to screen pools of T-DNA insertion lines, and isolated a mutant with an insertion in the *NPL1* gene (see Supplementary Information Fig. 2). Northern blot analysis revealed no *NPL1* mRNA in the mutant, implying that *npl1-1* is likely to be a loss-of-function, null allele (see Supplementary Information Fig. 2).

In view of the sequence similarity between NPH1 and NPL1, it seemed likely that the two proteins perform a similar function. As the *nph1* mutant exhibits altered phototropic behaviour<sup>12</sup>, we considered it likely that the *npl1* mutant could also be affected in this response. However, we observed little difference in the extent of phototropic curvature between the *npl1* mutant and wild-type seedlings at different fluence rates of blue light (data not shown).

Floral initiation is also regulated by blue light<sup>13</sup>. We investigated whether flowering time was affected in the *npl1-1* mutant. We found no differences in flowering time, in either long-day or short-day

conditions, between the mutant and corresponding wild-type plants (data not shown). Furthermore, the morphology and the growth rate of the *npl1-1* mutant, under a variety of light conditions, were identical to those of wild-type plants (data not shown).

For *Arabidopsis*, the *cry1cry2* and *nph1* mutants show normal blue-light-induced chloroplast relocation, although for the *nph1* mutant the sensitivity to low fluence rate light is slightly lower than that of the wild type<sup>14</sup>. We compared chloroplast distribution in the *npl1-1* mutant with that in Columbia wild-type plants by examining changes in red light transmittance through the leaves<sup>15</sup>. Leaves from dark-adapted *npl1* mutant plants showed the same transmittance (around 8%) as wild-type leaves (Fig. 2a). We examined chloroplast movement by determining the light transmittance of leaves irradiated with successively increasing fluence rates of blue light, with each light treatment lasting around 1 h. For wild-type plants, light transmittance decreased with increasing fluence rates up to  $4 \mu\text{mol m}^{-2} \text{s}^{-1}$  (Fig. 2a). This decrease in transmittance corresponds



**Figure 2** Chloroplast movement in response to continuous blue light. Changes in light transmission accompanying the relocation of chloroplasts were recorded photometrically. **a**, Fluence rate response curve. Dark-adapted leaves of *A. thaliana* with initial transmission of 8% were exposed to continuous blue light of fluence rates 0.08–120  $\mu\text{mol m}^{-2} \text{s}^{-1}$ . Each irradiation step lasted ~1 h. **b**, Light microscopy showing chloroplast movement. The microscope pictures were taken in wild-type (I and II) and *npl1*

mutant (III and IV) plants, at the times indicated in **a** after the chloroplasts had attained extreme positions at 4 and 120  $\mu\text{mol m}^{-2} \text{s}^{-1}$ . **c**, Responses of chloroplasts to weak, (4  $\mu\text{mol m}^{-2} \text{s}^{-1}$ ) intermediate (20  $\mu\text{mol m}^{-2} \text{s}^{-1}$ ) and strong (80  $\mu\text{mol m}^{-2} \text{s}^{-1}$ ) blue light. Changes in transmission were recorded in leaves of two dark-adapted *npl1* and wild-type plants. The initial transmission of the dark-adapted leaves was 9.5–10.0%. Fluence rates are indicated and the arrows denote the time at which the light was switched on.



to movement of chloroplasts from a random distribution around the cell to the periclinal walls, perpendicular to the incident light<sup>1,15</sup>. With greater fluence rates (20 and 40  $\mu\text{mol m}^{-2} \text{s}^{-1}$ ), the transmission showed a large increase, whereas irradiation with 120  $\mu\text{mol m}^{-2} \text{s}^{-1}$  resulted in little further change. We used light microscopy to show that these blue-light-induced changes in transmission were correlated with chloroplast movement. In leaves that had been irradiated with blue light of low fluence rate (4  $\mu\text{mol m}^{-2} \text{s}^{-1}$ ), we observed preferential distribution of chloroplasts at the periclinal walls (Fig. 2b). In contrast, for the leaves irradiated with high fluence rate light (120  $\mu\text{mol m}^{-2} \text{s}^{-1}$ ), the chloroplasts were primarily associated with the anticlinal walls.

When leaves from the mutant *npl1* plants were irradiated with low fluence rates of blue light, their transmittance decreased, in a manner not markedly different from that observed for leaves from wild-type plants (Fig. 2a). However, when irradiated with blue light of fluence rates of 20  $\mu\text{mol m}^{-2} \text{s}^{-1}$  or greater, the leaves from the mutant failed to show the increase in transmittance observed in wild-type leaves. These differences in transmission properties were correlated with a marked change in chloroplast movement: after irradiation with high fluence rates of blue light (120  $\mu\text{mol m}^{-2} \text{s}^{-1}$ ), the chloroplasts of the mutant retained their preferential association with the periclinal walls, similar to that observed under low fluence rates of light (Fig. 2b). These observations, along with the decrease in transmission observed when the mutant leaves were irradiated with low fluence rates of light, indicate that the *npl1* mutant shows altered chloroplast movement in response to high fluence rates of blue light, but behaves essentially like the wild type in response to low fluence rates of light.

The difference in chloroplast movement observed between the *npl1* mutant and wild-type plants was observed in an experiment where we monitored the light transmittance properties of leaves during continuous irradiation with different fluence rates of blue light (Fig. 2c). There were no detectable differences between the mutant and wild type when irradiation was with low fluence rates of light (4  $\mu\text{mol m}^{-2} \text{s}^{-1}$ ). Again, however, when the leaves were irradiated with high fluence rate light (20 or 80  $\mu\text{mol m}^{-2} \text{s}^{-1}$ ), the increase in transmission detected for the wild-type leaves was not observed for the mutant (Fig. 2c).

The striking difference in chloroplast movement in *npl1* showed that the avoidance response is lacking in the mutant. It follows from these observations, as well as the similarity in sequence between NPL1 and phototropin, that NPL1 is probably a photoreceptor associated with the movement of chloroplasts in response to high-fluence blue light. By contrast, NPL1 is not essential in the movement of chloroplasts in response to low fluence rates of light. These findings are consistent with earlier indications that there are two distinct mechanisms controlling chloroplast relocation, one functioning for the response to weak light and a second for the response to strong light. This was inferred from both the different effects of calcium depletion on the accumulation and avoidance responses<sup>16</sup> and a kinetic analysis of light-induced chloroplast movement<sup>17</sup>. In view of our findings, it will be of interest to evaluate chloroplast movement in response to low fluence rate light in the double *npl1 nph1* mutant. Whether NPL1 is involved in the phototropic response may also be apparent from a study of these double mutants. □

## Methods

### Detection of mRNA by northern hybridization analysis

We prepared root RNA from 3-week-old roots grown in liquid culture (Gamborg's B5 medium, Sigma) under white light. We prepared stem, leaf and flower RNA from 4-week-old plants that were grown in soil under white fluorescent bulbs (200  $\mu\text{mol m}^{-2} \text{s}^{-1}$ ). For light treatments, total RNA was prepared from 6-day-old etiolated seedlings, grown in Petri dishes containing Murashige and Skoog media (Sigma), and subsequently exposed for 6 h to ultraviolet-A (20  $\mu\text{mol m}^{-2} \text{s}^{-1}$ ), blue (20  $\mu\text{mol m}^{-2} \text{s}^{-1}$ ), red (15  $\mu\text{mol m}^{-2} \text{s}^{-1}$ ) and white light (75  $\mu\text{mol m}^{-2} \text{s}^{-1}$ ). The ultraviolet-B treatment was for 2 h and the lids of the Petri dishes were removed during irradiation. Light sources and filters were as described<sup>18</sup>. In all cases, RNA was electrophoresed on agarose gels and transferred to

Hybond N+ (Amersham). We analysed *NPL1* gene expression with a *NPL1* complementary DNA probe. The probe to detect *CAB1* gene expression was as described<sup>19</sup>. We carried out hybridization experiments following standard protocols. To evaluate the integrity and amount of RNA loaded in each lane of the RNA blots, filters were hybridized with an 18S ribosomal RNA probe. Experiments were repeated twice.

### Photometric measurement of chloroplast movements

We grew plants at 22 °C in a growth chamber, in a commercial mixture of peat moss and Perlite enriched in mineral nutrients. Illumination of 150  $\mu\text{mol m}^{-2} \text{s}^{-1}$  was provided by fluorescent lamps FL-40W (Polam), with a photoperiod of 15 h light/9 h dark. We used detached, fully developed leaves from plants between 2 and 4 weeks old for experiments. We measured movements of chloroplasts using a double-beam photometer<sup>20</sup>. In this system, chloroplast relocations are detected as changes in light transmission through the leaf tissue. The measuring monochromatic red light (660 nm, 0.3  $\mu\text{mol m}^{-2} \text{s}^{-1}$ ) has no influence on blue light-induced chloroplast responses and may be treated as physiological darkness. Before measurements, plants were adapted in darkness for at least 12 h. We used detached leaves with similar levels of transmission after dark adaptation for experiments. During the measurement, leaves were supplied with water through petioles and covered with saran-wrap to make respiration possible and to prevent loss of moisture. The temperature of adaptation and measurements was 22 ± 0.5 °C. The actinic light was obtained from a halogen lamp (100 W, 12 V) and filtered through a combination of filters (Schott, Jena): BG12, BG23, GG13, each 2 mm thick, and a heat-absorbing filter, C805 (3 mm). The resulting broad band is symmetrical, with the maximum at 430 nm and a half-band width of 80 nm. We used neutral density filters to reduce fluence rates to appropriate values. To take photographs, we removed the lower epidermis from the leaves and introduced water into intracellular spaces under weak negative pressure. These infiltrated leaf fragments were placed on a slide in a drop of water and irradiated in the photometer with the simultaneous measurement of transmission level. Photographs were taken after transmission reached a stationary minimum or maximum level.

Received 27 December 2000; accepted 8 March 2001.

- Zurzycki, J. In *Blue-light Syndrome* (ed. Senger, H.) 50–68 (Springer, Berlin, 1980).
- Jarillo, J. A., Ahmad, M. & Cashmore, A. R. NPL1 (Accession number AF053941): a second member of the NPH serine/threonine kinase family of Arabidopsis. *Plant Physiol.* **117**, 719 (1998).
- Huala, E. et al. Arabidopsis NPH1: a protein kinase with a putative redox-sensing domain. *Science* **278**, 2120–2123 (1997).
- Wada, M., Grolig, F. & Haupt, W. Light-oriented chloroplast positioning. Contribution to progress in photobiology. *J. Photochem. Photobiol.* **17**, 3–25 (1993).
- Kadota, A., Sato, Y. & Wada, M. Intracellular chloroplast photorelocation in the moss *Physcomitrella patens* is mediated by phytochrome as well as a blue-light receptor. *Planta* **210**, 932–937 (2000).
- Briggs, W. R. & Huala, E. Blue-light photoreceptors in higher plants. *Annu. Rev. Cell Dev. Biol.* **15**, 33–62 (1999).
- Zacherl, M., Huala, E., Rudiger, W., Briggs, W. R. & Salomon, M. Isolation and characterization of cDNAs from oat encoding a serine/threonine kinase: an early component in signal transduction for phototropism. *Plant Physiol.* **116**, 869 (1998).
- Kanegae, H. et al. Rice NPH1 homologues, OsNPH1a and OsNPH1b, are differently photoregulated. *Plant Cell Physiol.* **41**, 415–423 (2000).
- Noze, K., Christie, J. M., Kiyosue, T., Briggs, W. R. & Wada, M. Isolation and characterization of a fern phototropin (Accession number AB037188), a putative blue-light photoreceptor for phototropism. *Plant Physiol.* **122**, 1457 (2000).
- Salomon, M., Christie, J. M., Knieb, E., Lempert, U. & Briggs, W. R. Photochemical and mutational analysis of the FMN-binding domains of the plant blue light photoreceptor, phototropin. *Biochemistry* **39**, 9401–9410 (2000).
- Hanks, S. K., Quinn, A. M. & Hunter, T. The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science* **241**, 42–52 (1988).
- Liscum, E. & Briggs, W. R. Mutations in the NPH1 locus of Arabidopsis disrupt the perception of phototropic stimuli. *Plant Cell* **7**, 473–485 (1995).
- Lin, C. Photoreceptors and regulation of flowering time. *Plant Physiol.* **123**, 39–50 (2000).
- Kagawa, T. & Wada, M. Blue light-induced chloroplast relocation in *Arabidopsis thaliana* as analyzed by microbeam irradiation. *Plant Cell Physiol.* **41**, 84–93 (2000).
- Trojan, A. & Gabrys, H. Chloroplast distribution in *Arabidopsis thaliana* (L.) depends on light conditions during growth. *Plant Physiol.* **111**, 419–425 (1996).
- Tlalka, M. & Gabrys, H. Influence of calcium on blue-light-induced chloroplast movement in *Lemna trisulca*. *Planta* **189**, 491–498 (1993).
- Malec, P. Kinetic modelling of chloroplast phototranslocations in *Lemna trisulca* L: two rate-limiting components? *J. Theor. Biol.* **169**, 189–195 (1994).
- Ahmad, M. & Cashmore, A. R. HY4 gene of *A. thaliana* encodes a protein with characteristics of a blue-light photoreceptor. *Nature* **366**, 162–166 (1993).
- Leutwiler, L. S., Meyerowitz, E. & Tobin, E. M. Structure and expression of three light-harvesting chlorophyll a/b binding protein genes of *Arabidopsis thaliana*. *Nucleic Acids Res.* **14**, 4051–4064 (1986).
- Walczak, T. & Gabrys, H. New type of photometer for measurements of transmission changes corresponding to chloroplast movements in leaves. *Photosynthetica* **14**, 65–72 (1980).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

### Acknowledgements

This work was supported by grants to A.R.C. from NIH and DOE, and a NATO fellowship awarded to J.C. We thank N. Bonini and S. Poethig for comments on the manuscript.

Correspondence and requests for materials should be addressed to A.R.C. (e-mail: cashmore@sas.upenn.edu).

# Degradation of a cohesin subunit by the N-end rule pathway is essential for chromosome stability

Hai Rao\*, Frank Uhlmann†, Kim Nasmyth‡ & Alexander Varshavsky\*

\* Division of Biology, California Institute of Technology, Pasadena, California 91125, USA

† Imperial Cancer Research Fund, 44 Lincoln's Inn Fields, London WC2A 3PX, UK

‡ Research Institute of Molecular Pathology, Dr Bohr Gasse 7, A-1030 Vienna, Austria

Cohesion between sister chromatids is established during DNA replication and depends on a protein complex called cohesin<sup>1–7</sup>. At the metaphase–anaphase transition in the yeast *Saccharomyces cerevisiae*, the ESP1-encoded protease separin cleaves SCC1, a subunit of cohesin with a relative molecular mass of 63,000 (M<sub>r</sub> 63K)<sup>8</sup>. The resulting 33K carboxy-terminal fragment of SCC1 bears an amino-terminal arginine—a destabilizing residue in the N-end rule<sup>9</sup>. Here we show that the SCC1 fragment is short-lived ( $t_{1/2} \approx 2$  min), being degraded by the ubiquitin/proteasome-dependent N-end rule pathway. Overexpression of a long-lived derivative of the SCC1 fragment is lethal. In *ubr1Δ* cells, which lack the N-end rule pathway<sup>9</sup>, we found a highly increased frequency of chromosome loss. The bulk of increased chromosome loss in *ubr1Δ* cells is caused by metabolic stabilization of the ESP1-produced SCC1 fragment. This fragment is the first physiological substrate of the N-end rule pathway that is targeted through its N-terminal residue. A number of yeast proteins bear putative cleavage sites for the ESP1 separin, suggesting other physiological substrates and functions of the N-end rule pathway.

The sister chromatids of a replicated chromosome are pulled apart by the centromere-attached microtubules that emanate from spindle poles<sup>1–6</sup>. However, until the onset of anaphase the forces pulling sister chromatids apart are counteracted by cohesion that holds sisters together. A complex termed cohesin is essential for sister chromatid cohesion in both fungi and metazoans. In the yeast *S. cerevisiae*, cohesin contains four subunits, SMC1, SMC3, SCC1 (also called MCD1 and RAD21) and SCC3 (see refs 1–5). Perturbations of chromosome segregation are among the causes of human cancer and other diseases.

Previous work has shown that *S. cerevisiae* SCC1, a 63K subunit of cohesin, is cleaved at the metaphase–anaphase transition, in a reaction that requires the 187K ESP1, termed separin<sup>8</sup>. ESP1, a cysteine protease of a class that includes caspases<sup>10</sup>, is activated at the onset of anaphase through ubiquitin (Ub)-dependent destruction of its inhibitory ligand PDS1 (see refs 1–3, 7). A protein substrate of the Ub system is conjugated to Ub through the action of enzymes E1, E2 and E3, with the degradation signal (degron) of the substrate recognized by E3. The resulting multiubiquitylated substrate is degraded by the 26S proteasome (see refs 11, 12).

The ESP1 separin cleaves the 566-residue SCC1 at two sites, 180 and 268 residues from its N terminus, the latter cleavage site being the principal one<sup>8</sup> (Fig. 1a). Mutations in both (but not single) SCC1 cleavage sites that inhibit the ESP1-mediated cleavage of SCC1 are lethal; they preclude sister chromatid separation<sup>8,10</sup>. The N-terminal Arg of ESP1-produced SCC1 fragments is a destabilizing residue in the N-end rule<sup>9,13</sup>. The degrons recognized by the N-end rule pathway, called N-degrons, consist of two determinants, a destabilizing N-terminal residue and an internal Lys residue<sup>14</sup>. UBR1, a 225K, RING-H2 finger-containing E3 of the N-end rule pathway<sup>15</sup>, recognizes eight primary destabilizing N-terminal residues: Arg, Lys, His, Phe, Leu, Tyr, Trp and Ile. The other four

destabilizing residues, Asp and Glu (secondary), and Asn and Gln (tertiary), function through their enzymatic modification and conjugation to Arg, a primary destabilizing residue<sup>9</sup>. UBR1 has at least three substrate-binding sites. The type 1 site is specific for N-terminal Arg, Lys and His. The type 2 site is specific for N-terminal Phe, Leu, Trp, Tyr and Ile. The third site of UBR1 is specific for proteins such as CUP9 and GPA1, which bear internal (non-N-terminal) degrons<sup>16</sup>. The functions of the N-end rule pathway include the control of peptide import in *S. cerevisiae*, by regulated degradation of CUP9—a transcriptional repressor of the PTR2-encoded peptide transporter<sup>16</sup>. Engineered N-end rule substrates can be produced through the Ub fusion technique, in which a Ub-reporter fusion is cleaved, co-translationally, at the Ub-reporter junction by deubiquitylating enzymes (DUBs), yielding a reporter bearing a predetermined N-terminal residue<sup>13,17</sup>. Neither type 1 nor type 2 physiological N-end rule substrates have yet been identified in *S. cerevisiae*.

To determine whether the ESP1-produced SCC1 fragments<sup>8</sup> were degraded by the N-end rule pathway, we took advantage of the fact that dipeptides bearing destabilizing N-terminal residues can specifically inhibit the yeast N-end rule pathway *in vivo*, through their import into the cell and interaction with the type 1 or type 2 sites of UBR1 (ref. 16). *S. cerevisiae* K6843, in which the full-length, C-terminally Myc<sub>18</sub>-epitope-tagged SCC1 (termed SCC1<sup>m</sup>) was expressed from the P<sub>GAL1-10</sub> promoter<sup>8</sup>, were arrested in a metaphase-like state with nocodazole, followed by the induction of SCC1<sup>m</sup> with galactose. The dipeptides Ala-Arg, Arg-Ala, Trp-Ala or Leu-Ala were added 30 min later, and the incubation was continued for another 30 min, followed by extraction of proteins and anti-Myc immunoblotting.

In the absence of added dipeptides, we could detect only uncleaved SCC1<sup>m</sup> (Fig. 1b, lane 2). However, in the presence of Arg-Ala, bearing a type 1 destabilizing N-terminal residue, an additional, smaller species was observed, whose size was consistent with the ESP1-produced major C-terminal fragment of SCC1<sup>m</sup>, termed SCC1<sup>269–566m</sup> (Fig. 1b, lane 4). This band was absent from cells treated either with Ala-Arg, which bore a stabilizing N-terminal residue, or with Trp-Ala or Leu-Ala, both of which bore a type 2 destabilizing N-terminal residue (Fig. 1b, lanes 1–6). Crucially, SCC1<sup>269–566m</sup> could also be observed in the absence of dipeptides, provided that the cells lacked *UBR1* and therefore could not degrade N-end rule substrates (Fig. 1b, lanes 4, 7). No SCC1 fragment derived from cleavage solely at the minor site (residues 180–181) was observed (Fig. 1b). We conclude, in agreement with an earlier inference<sup>8</sup>, that the ESP1-mediated minor-site cleavage of SCC1 is either negligible or rapidly followed by cleavage at the major site.

To determine the *in vivo* half-life of Arg-SCC1<sup>269–566</sup>, it was expressed as part of a fusion of the form <sup>1</sup>DHFR–Ub–Arg-SCC1<sup>269–566f</sup>, where <sup>1</sup> denotes the N-terminal and C-terminal Flag epitopes linked, respectively, to mouse dihydrofolate reductase (DHFR) and Arg-SCC1<sup>269–566</sup> (Fig. 1a). DUBs co-translationally cleave this fusion at the C terminus of the Ub moiety, yielding the test protein Arg-SCC1<sup>269–566f</sup> and the long-lived <sup>1</sup>DHFR–Ub reference protein, which serves as an internal control<sup>14,16,18</sup>. In agreement with dipeptide inhibition experiments (Fig. 1b), Arg-SCC1<sup>269–566f</sup> was short-lived in *UBR1* cells ( $t_{1/2} \approx 2$  min) (Fig. 1c, lanes 1–4). By contrast, the same Arg-SCC1<sup>269–566f</sup> protein was long-lived ( $t_{1/2} > 2$  h) in *ubr1Δ* cells, indicating that its degradation in wild-type cells was mediated by the N-end rule pathway (Fig. 1c, lanes 5–8; and 1e). Arg-SCC1<sup>269–566f</sup> was destroyed particularly rapidly either during or shortly after its synthesis, that is, during the pulse. Specifically, nearly 90% of Arg-SCC1<sup>269–566f</sup> was degraded by the N-end rule pathway during the 3-min pulse; the rest of pulse-labelled Arg-SCC1<sup>269–566f</sup> was destroyed more slowly ( $t_{1/2} \approx 2$  min) (Fig. 1e). If the N-terminal Arg of Arg-SCC1<sup>269–566f</sup> was converted to Met, a stabilizing residue in the N-end rule, the resulting Met-

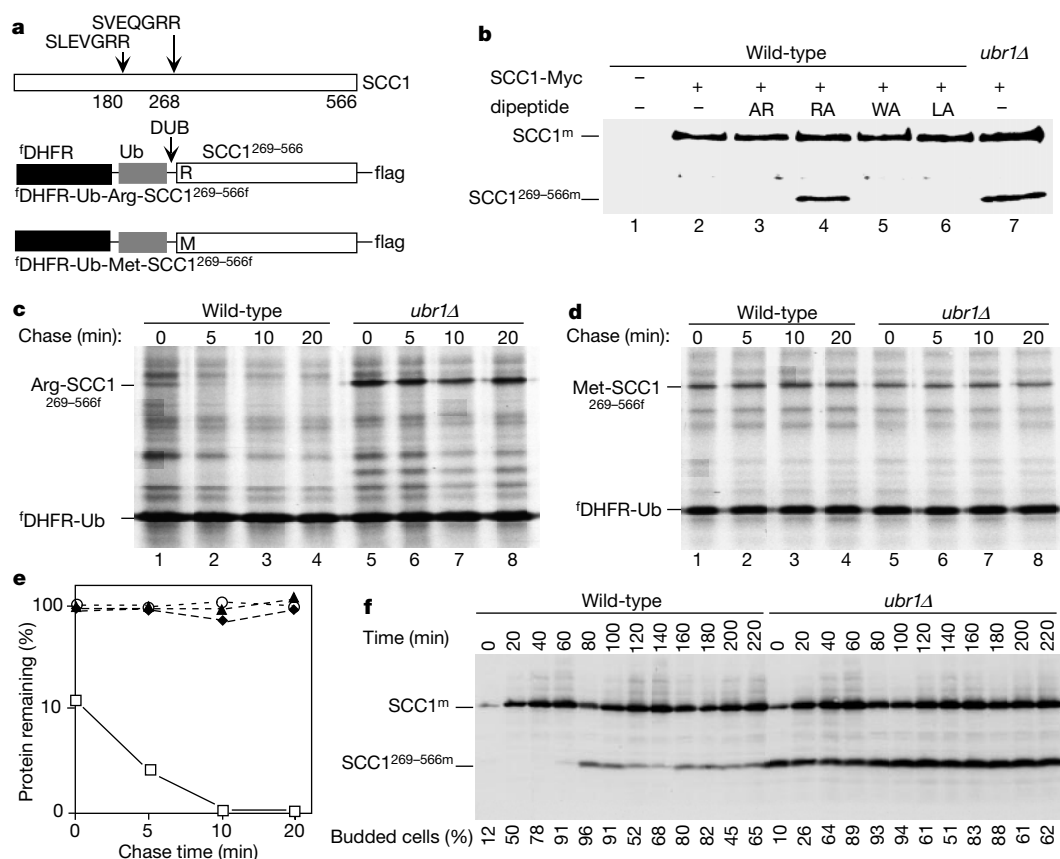
SCC1<sup>269–566f</sup> (Fig. 1a) was long-lived in both *UBR1* and *ubr1Δ* cells, in contrast to Arg-SCC1<sup>269–566f</sup> (Fig. 1c, d). We conclude that Arg-SCC1<sup>269–566f</sup> is targeted for degradation through its Arg-bearing N-degron, rather than through an internal (non-N-terminal) degradation signal.

Overexpression of the N-terminal SCC1 fragment (residues 1–268; Fig. 1a) in either *UBR1* or *ubr1Δ* cells did not significantly alter cell growth (data not shown). In contrast, the analogous expression of Arg-SCC1<sup>269–566f</sup> was lethal in *ubr1Δ* cells (Fig. 2a). In *UBR1* cells, where Arg-SCC1<sup>269–566f</sup> was short-lived (Fig. 2a), its overexpression was not lethal but slowed growth (data not shown). Overexpression of the Met-SCC1<sup>269–566f</sup> was lethal in both *UBR1* and *ubr1Δ* cells, consistent with metabolic stability of this SCC1 derivative in both genetic backgrounds (Fig. 2a).

To examine toxicity of the long-lived Met-SCC1<sup>269–566f</sup> in more detail, we transiently overexpressed Met-SCC1<sup>269–566f</sup> from the *P<sub>GALI</sub>* promoter in cells that were arrested either before the S phase (with  $\alpha$ -factor), or in the S phase (with hydroxyurea), or in a metaphase-like state (with nocodazole). Transient overexpression of Met-

SCC1<sup>269–566f</sup> was largely nontoxic to cells in G1 and S phases but highly toxic to mitotic (nocodazole-arrested) cells (Fig. 2b), indicating that elevated levels of SCC1<sup>269–566f</sup> may interfere with a process that takes place specifically at metaphase–anaphase. Control immunoblots confirmed that the levels of Met-SCC1<sup>269–566f</sup> were about equal in cells arrested at either of the three phases of the cell cycle (data not shown). Furthermore, we found that cells remained stably arrested in the presence of nocodazole even after overexpression of the SCC1 fragment (data not shown), thus making it unlikely that abnormally high levels of this fragment abrogated the MAD2-dependent checkpoint and the cell-cycle arrest induced by nocodazole. It is also possible that selective toxicity of the SCC1 fragment in nocodazole-arrested cells is caused by this fragment acting as a competitive inhibitor of the ESP1 protease.

Full-length SCC1 interacts with SMC1, a DNA-binding subunit of cohesin<sup>1,2</sup>. Therefore one cause of lethality of overexpressed SCC1<sup>269–566f</sup> may be a residual (and functionally perturbing) affinity of this fragment for the rest of the cohesin complex. In this model,



**Figure 1** The SCC1 fragment is a substrate of the N-end rule pathway. **a**, The *in vivo* cleavage sites in *S. cerevisiae* SCC1 (ref. 8), and the UPR-type fusions that produce, on co-translational cleavage by DUBs, the Flag-tagged reference protein <sup>1</sup>DHFR–Ub and the Flag-tagged test proteins Arg-SCC1<sup>269–566f</sup> or Met-SCC1<sup>269–566f</sup>. **b**, Detection of SCC1 fragment. Control *S. cerevisiae* cells and cells expressing SCC1<sup>m</sup> were incubated with specific dipeptides, followed by immunoblotting with anti-Myc antibody. Lane 1, control cells (lacking SCC1<sup>m</sup>). Lane 2, cells expressing SCC1<sup>m</sup> in the absence of added dipeptide. Lanes 3–6, as lane 2 but in the presence of Ala-Arg (AR), Arg-Ala (RA), Trp-Ala (WA) and Leu-Ala (LA), respectively. Lane 7, *ubr1Δ* cells expressing SCC1<sup>m</sup> in the absence of dipeptides. Arrows indicate bands of SCC1<sup>m</sup> and SCC1<sup>269–566m</sup>. **c**, Pulse–chase analysis of Arg-SCC1<sup>269–566f</sup>. Cells expressing <sup>1</sup>DHFR–Ub–Arg-SCC1<sup>269–566f</sup> were labelled for 3 min with [<sup>35</sup>S]methionine, followed by a chase for 5, 10 or 20 min, and SDS–PAGE analysis of anti-Flag immunoprecipitates. Lanes 1–4, pulse chase in *UBR1* cells. Lanes 5–8, pulse chase in *ubr1Δ* cells. Arrows indicate bands of Arg-SCC1<sup>269–566f</sup> and the

reference protein <sup>1</sup>DHFR–Ub. **d**, As in **c**, but with cells expressing Met-SCC1<sup>269–566f</sup>, derived from <sup>1</sup>DHFR–Ub–Met-SCC1<sup>269–566f</sup>. **e**, Quantification of results in **c** and **d** by PhosphorImager. Squares, Arg-SCC1<sup>269–566f</sup> in JD52 (*UBR1*) cells; diamonds, Arg-SCC1<sup>269–566f</sup> in JD55 (*ubr1Δ*) cells; circles, Met-SCC1<sup>269–566f</sup> in JD52 (*UBR1*) cells; triangles, Met-SCC1<sup>269–566f</sup> in JD55 (*ubr1Δ*) cells. The amounts of <sup>35</sup>S in the test proteins, relative to <sup>35</sup>S in the <sup>1</sup>DHFR–Ub reference protein, were plotted as percentages of this ratio for Met-SCC1<sup>269–566f</sup> in *UBR1* cells at time zero. **f**, Immunoblot analysis, using anti-Myc antibody, of Arg-SCC1<sup>269–566m</sup> produced through ESP1-mediated cleavage of normally expressed, C-terminally Myc-tagged SCC1<sup>m</sup> in wild-type (*UBR1*) and *ubr1Δ* cultures synchronized with  $\alpha$ -factor, as described<sup>8</sup>. The strains used were *S. cerevisiae* Y92 (*MATa ScclMyc18<sup>+</sup>*) and Y238 (*MATa ScclMyc18 ubr1Δ*). Shown are the time after removal of  $\alpha$ -factor (top) and the content of budded (post-G1 phase) cells (bottom) in *UBR1* and congenic *ubr1Δ* cultures.



the function of metabolic instability of the SCC1<sup>269–566</sup> fragment is to ensure its removal from the cohesin complex, allowing re-binding of intact SCC1 in the next cell cycle. To determine whether Met-SCC1<sup>269–566f</sup> could, in fact, interact with SMC1, extracts from cells expressing Flag-tagged Met-SCC1<sup>269–566f</sup>, either alone or together with RGS/His<sub>6</sub>-tagged SMC1, were precipitated with anti-Flag antibody, followed by SDS polyacrylamide gel electrophoresis (SDS–PAGE) of the immunoprecipitates and immunoblotting with either anti-Flag or anti-RGS/His<sub>6</sub> antibodies. We found that Met-SCC1<sup>269–566f</sup> was specifically co-immunoprecipitated with the SMC1 (Fig. 2c). However, overexpression of SMC1, from the P<sub>GALI</sub> promoter and low copy plasmid, did not suppress the lethal effect of Met-SCC1<sup>269–566f</sup> overexpression in *UBR1* cells (data not shown). Thus, the toxicity of overexpressed SCC1<sup>269–566</sup> may not be caused exclusively by sequestration of cohesin complexes.

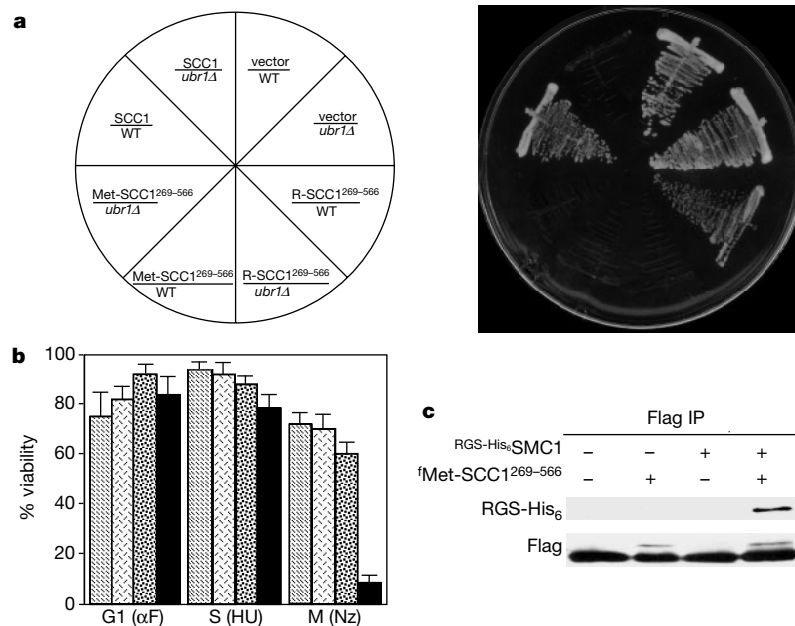
Given the above results (Figs 1 and 2), might *ubr1Δ* cells, which lack the N-end rule pathway and in which the Arg-SCC1 fragment is long-lived, be impaired in the maintenance of chromosome stability? We used a colony sectoring assay<sup>19</sup> to compare the rates of chromosome loss in *ubr1Δ* and *UBR1* cells (Fig. 3a). Remarkably, whereas less than 0.5% of *UBR1* colonies contained red sectors, ~22% of the *ubr1Δ* colonies contained such sectors (Fig. 3a; and data not shown), revealing a previously unknown phenotype of chromosome instability in cells lacking the N-end rule pathway. The frequency of distinct red sectors (Fig. 3a) is roughly proportional to the frequency of cells that suffered the loss of a (non-essential) marked chromosome during the growth of colonies. We conclude that *ubr1Δ* cells lose chromosomes about 100 times more frequently than their congenic *UBR1* counterparts.

To determine how much of the chromosome-loss phenotype of *ubr1Δ* cells was caused by their inability to degrade specifically the

Arg-SCC1 fragment (as distinguished from other N-end rule substrates), we constructed a *S. cerevisiae* strain (in the background of a sectoring strain<sup>19</sup>) in which wild-type SCC1 was replaced by an allele (expressed from the native P<sub>SCC1</sub> promoter) that encoded SCC1 bearing at its C terminus three tandem haemagglutinin A (HA) tags and lacking the minor ESP1 cleavage site (Fig. 1a), owing to an Arg to Glu mutation at position 180. In this strain, the only ESP1-produced SCC1 fragment was Arg-SCC1<sup>269–566h</sup>. We also constructed a derivative of this strain in which Arg-269 of SCC1–HA was replaced with Gly. The resulting ESP1-produced fragment of SCC1 was the long-lived Gly-SCC1<sup>269–566h</sup> bearing a stabilizing N-terminal residue.

We found that the efficiency of cleavage at the thus altered cleavage site of SCC1 was indistinguishable from that at the wild-type cleavage site (Fig. 3b). In the strain producing the wild-type (short-lived) Arg-SCC1<sup>269–566</sup> fragment, the frequency of chromosome loss was indistinguishable from that in the congenic wild-type strain (Fig. 3a; and data not shown). By contrast, in the strain in which the ESP1-produced Gly-SCC1<sup>269–566h</sup> fragment bore a stabilizing N-terminal residue, a greatly increased frequency of chromosome loss was observed, similarly to congenic *ubr1Δ* cells expressing wild-type SCC1 (Fig. 3a; and data not shown). We conclude that the bulk of increased chromosome loss in *ubr1Δ* cells is caused by metabolic stabilization of the ESP1-produced SCC1 fragment.

Although the steady-state level of ESP1-generated Arg-SCC1<sup>269–566h</sup> was significantly (~3-fold) lower in wild-type (*UBR1*) than in *ubr1Δ* cells (Fig. 3b, lanes 2 and 3), the observed difference was less than would be expected from the extremely short (~2 min) half-life of overproduced Arg-SCC1<sup>269–566</sup> in *UBR1* cells (Fig. 1c, d). To assess the influence of the N-end rule pathway on the concentration of endogenous, ESP1-produced Arg-SCC1<sup>269–566</sup> fragment, we monitored its levels by immunoblotting as a function of time,



**Figure 2** Effects of overexpressing SCC1 and its fragments in *UBR1* and *ubr1Δ* cells. **a**, Overexpression of long-lived SCC1<sup>269–566</sup> is lethal. Strains K6843 (*UBR1*), JD52 (*UBR1*) or JD55 (*ubr1Δ*) carrying plasmids that overexpressed, from the P<sub>GALI</sub> promoter, full-length SCC1, Arg-SCC1<sup>269–566</sup> or Met-SCC1<sup>269–566f</sup> (left) were streaked either on dextrose-containing (SD) plates (not shown), on which all strains grew, or on galactose-containing (SG) plates (right). The plates were incubated for 3 d at 30 °C. **b**, Plating efficiency of *UBR1* cells that transiently expressed Met-SCC1<sup>269–566</sup> while being transiently arrested in late G1 phase (with α-factor, αF), in S phase (with hydroxyurea, HU), or in a metaphase-like state (with nocodazole, Nz). For each set of assays: the first bar

refers to wild-type (*UBR1*) AVY110 cells incubated in raffinose-containing medium (YP-R); the second to the same strain in galactose-containing YP medium (YP-G); the third to congenic AVY111 cells in YP-R that carried the ORF encoding Met-SCC1<sup>269–566f</sup> downstream from the P<sub>GALI</sub> promoter; and the fourth to the same strain in YP-G. **c**, The SCC1<sup>269–566</sup> fragment interacts with SMC1. Extracts of *S. cerevisiae* expressing either Met-SCC1<sup>269–566f</sup> (Flag-tagged), or RGS/His<sub>6</sub>-tagged SMC1, or both were precipitated with anti-Flag antibody, followed by SDS–PAGE and immunoblotting with either anti-Flag or anti-RGS/His<sub>6</sub>. Asterisk denotes a protein crossreacting with anti-Flag antibody.

using synchronized cultures of *UBR1* and *ubr1Δ* cells. The levels of full-length SCC1 were regulated similarly in *UBR1* and *ubr1Δ* cells (Fig. 1f). In a *UBR1* culture, cells in G1 phase contained virtually no Arg-SCC1<sup>269–566</sup> fragment, which appeared transiently as cells entered anaphase (Fig. 1f). In striking contrast, *ubr1Δ* cells contained the Arg-SCC1<sup>269–566</sup> fragment throughout the cell cycle, and in addition at levels that significantly exceeded the peak level of Arg-SCC1<sup>269–566</sup> in *UBR1* cells (Fig. 1f).

Together, the short-half life of the Arg-SCC1<sup>269–566</sup> fragment (Fig. 1e), the toxicity of its long-lived counterpart (Fig. 2), the altered accumulation pattern and steady-state levels of Arg-SCC1<sup>269–566</sup> in *ubr1Δ* cells (Fig. 1f), the greatly increased chromosome loss in *ubr1Δ* cells (Fig. 3a), and the evidence that most of this loss is caused by metabolic stabilization of Arg-SCC1<sup>269–566</sup> allow the following conclusions. First, if the Arg-SCC1<sup>269–566</sup> fragment becomes long-lived (as it does in *ubr1Δ* cells), even the relatively low, wild-type level of its expression is detrimental to the fidelity of chromosome maintenance. Second, the metabolic stabilization of N-end rule substrates other than Arg-SCC1<sup>269–566</sup> seems not to

contribute significantly to the elevated chromosome loss in *ubr1Δ* cells. As to the cause of toxicity of overexpressed SCC1<sup>269–566</sup> fragment, one possibility, suggested by high SCC1<sup>269–566</sup> levels in G1-phase *ubr1Δ* cells (Fig. 1f), is that the fragment interferes with re-formation of cohesin complexes that are required for the establishment of cohesion during the S phase.

The increased probability of chromosome loss in *ubr1Δ* cells, although readily detectable through the sectoring assay (Fig. 3a), is still below levels that would cause massive cell death during growth of a culture. In particular, no significant differences were observed between congenic *ubr1Δ* and *UBR1* cells in the kinetics of cell-cycle progression (data not shown). However, a further significant increase in the level of long-lived SCC1<sup>269–566</sup> (achieved through its transcriptional overexpression) was found to be lethal (Fig. 2a). A parsimonious interpretation is that the latter effect is mechanistically similar to the much less severe, viability-compatible chromosome-loss phenotype of *ubr1Δ* cells, in which SCC1<sup>269–566</sup> becomes long-lived but is not overexpressed otherwise. In this interpretation, the frequency of chromosome loss in cells that overexpress long-lived SCC1<sup>269–566</sup> becomes high enough for them to lose a chromosome nearly every division. Light-microscopic observations indicated that individual cells divided once or twice after overexpression of Met-SCC1<sup>269–566</sup>, followed by growth arrest and a varying terminal phenotype (data not shown).

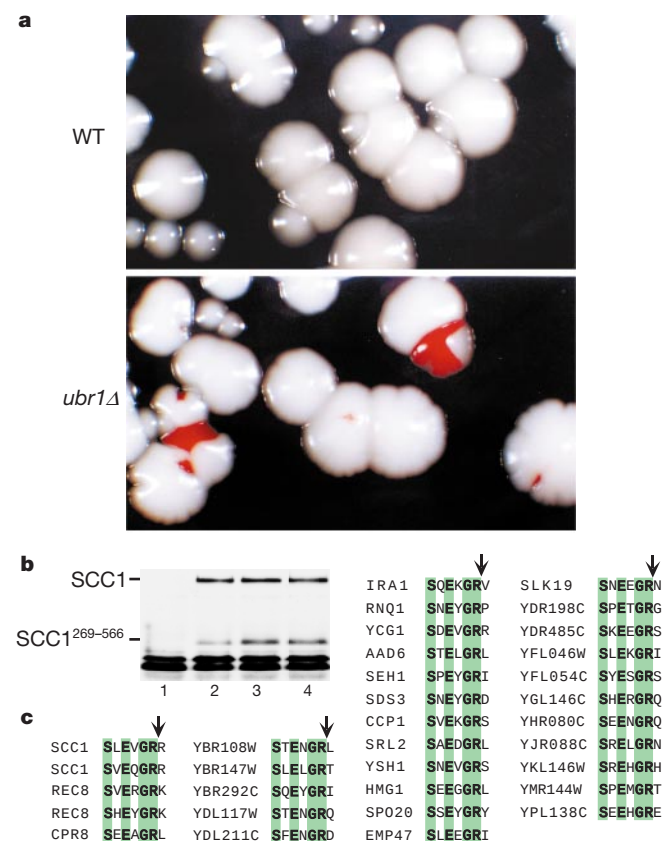
Some physiological N-end rule substrates other than SCC1 may also be produced by the ESP1 separin. The consensus sequence of the two ESP1 cleavage sites in SCC1 is SxExGR<sup>+</sup>R (Fig. 1a). *S. cerevisiae* REC8, the meiosis-specific counterpart of SCC1, contains two near-matches to this consensus<sup>8</sup>. REC8 is cleaved by ESP1 during yeast meiosis, similarly to SCC1 in vegetative cells<sup>20</sup>, and potential cleavage sites are also present in *Schizosaccharomyces pombe* RAD21 and REC8, the fission yeast homologues of SCC1 and REC8 (ref. 8). Simultaneous (but not single) alterations of these sites in *S. pombe* RAD21 block chromosome segregation, similarly to the results in *S. cerevisiae*<sup>21</sup>. A shorter consensus, SxExGR<sup>+</sup>, derived from comparisons of the ESP1 cleavage sites in the budding and fission yeast SCC1 homologues<sup>8</sup>, was used to search the *S. cerevisiae* genome database, yielding 29 putative substrates of ESP1 other than SCC1 and REC8 (Fig. 3c). The N-terminal residues of the predicted C-terminal fragments of these proteins encompass nearly the entire set of 12 residues that are destabilizing in the N-end rule, including tertiary and secondary destabilizing residues (Fig. 3c).

Until now, the known substrates of the N-end rule pathway that are recognized through an N-degron were either polypeptide-derived viral proteins<sup>9,22</sup> or proteins secreted by an intracytoplasmic bacterium such as *Listeria monocytogenes* into the mammalian cell's cytosol<sup>23</sup>. Processing proteases that convert pro-N-degrons into N-degrons have long been envisioned as regulated entry points to the N-end rule pathway<sup>9,13</sup>. The identification of ESP1 as one such entry point, and of SCC1 as the first N-degron-based physiological N-end rule substrate in yeast, marks the initial understanding of connections between the N-end rule and the ESP1 separin. The discovery of increased chromosome instability in *ubr1Δ* cells suggests that chromosome aneuploidy and related mutator phenotypes characteristic of mammalian cancer cells may also involve perturbations of the N-end rule pathway. Because ESP1 homologues are present in most if not all eukaryotes<sup>24</sup>, identifying ESP1 substrates in these organisms should be an effective way to pinpoint physiological N-end rule substrates, as well as additional, currently unknown, functions of the N-end rule pathway. □

## Methods

### Yeast strains and plasmids

Strains *S. cerevisiae* JD52 (*MATa lys2-801 ura3-52 trp1-Δ63 his3-Δ200 leu2-3,112*) and JD55 (*ubr1::HIS3* in the JD52 background) have been described<sup>25</sup>. *S. cerevisiae* K6843 (*MATa bar1 ade2-1 trp1-1 can1-100 leu2-3,112 his3-11,15 Gal-ScclMyc18::URA3*) expressed full-length, C-terminally Myc<sub>18</sub>-tagged SCC1 from the P<sub>GAL1-10</sub> promoter<sup>8</sup>. We



**Figure 3** Increased chromosome instability in *ubr1Δ* *S. cerevisiae*. **a**, Sectoring assay for the chromosome loss using the *SUP11*-marked YPH277 (*UBR1*) strain and its *ubr1Δ* derivative (see Methods). **b**, Immunoblot analysis of Arg-SCC1<sup>269–566h</sup> and Gly-SCC1<sup>269–566h</sup>, produced through ESP1-mediated cleavage of normally expressed, C-terminally HA-tagged SCC1 in wild-type and *ubr1Δ* strains. Yeast cultures ( $A_{600} \approx 0.4$ ) were processed for immunoblotting with anti-HA antibody (Covance, Berkeley, CA). Lane 1, YPH277 (control) cells lacking the HA-tagged SCC1. Lane 2, AVY129 (*UBR1*) cells expressing normal levels of SCC1<sup>R180E</sup>-HA. Lane 3, as lane 2 but with AVY132 (*ubr1Δ*) cells. Lane 4, as lane 2, but with SCC1<sup>R180E, R269G</sup>-HA, bearing the Arg to Gly mutation at position 269. The bands of full-length SCC1-HA and its ESP1-produced SCC1<sup>269–566h</sup> fragments are indicated on the left. **c**, *S. cerevisiae* proteins containing putative cleavage sites for the ESP1 separin. The query SxExGRx and the Pattern Match program (<http://genome-www2.stanford.edu/cgi-bin/SGD/PATMATCH/nph-patmatch>) were used to search the *S. cerevisiae* genome database. Arrows indicate the inferred sites of cleavage by ESP1.

produced AVY115 (*ubr1::LEU2*), a UBR1-lacking derivative of K6843, using standard techniques. The plasmids p426CUPR-SCC1, p426CUPR-SCC1R, p426CUPR-SCC1M and p426CUPR-SCC1N expressed, respectively, full-length SCC1<sup>f</sup>, Arg-SCC1<sup>269-566f</sup>, Met-SCC1<sup>269-566f</sup> and SCC1<sup>1-268f</sup> from the P<sub>CUP1</sub> promoter as UPR<sup>14</sup> fusions of the form <sup>f</sup>DHFR-Ub-(SCC1 derivative) (Fig. 1a). The Ub moiety of these fusions bore a Lys to Arg alteration at position 48.

<sup>f</sup>DHFR-Ub-Arg-SCC1<sup>269-566f</sup> and <sup>f</sup>DHFR-Ub-Met-SCC1<sup>269-566f</sup> were expressed from the P<sub>GAL1</sub> promoter in the plasmids p416UPRSCC1R and p416UPRSCC1M, respectively, which were derived from the low-copy vector p416Gal1 (ref. 26). The N-terminally RGS/His<sub>6</sub>-tagged SMC1 was expressed from the P<sub>GAL1</sub> promoter in plasmid p415GalSMC1, derived from p415Gal1 (ref. 26).

We cloned a fragment encoding Met-SCC1<sup>269-566f</sup> downstream from the P<sub>GAL1</sub> promoter into pRS305 (ref. 25). The resulting plasmid was cut with *XcmI* within LEU2 and transformed into *S. cerevisiae* AVY110 (*MATa bar1-1 ade2-1 leu2-3,112 ura3-1*), yielding AVY111 (*leu2::P<sub>GAL1</sub>-Met-SCC1<sup>269-566f</sup>::LEU2*). UBR1 was deleted from the sectoring-assay strain YPH277 (*MATa ura3-52 ade2-101 trp1-Δ1 leu2-Δ1 CFVII (RAD2.d) URA3 SUP11*)<sup>19</sup>, yielding YPH277HR1 (*ubr1::LEU2*). We also constructed, in the background of the pYIplac128 vector, the alleles of SCC1 encoding the triple-HA tag at the SCC1 C terminus<sup>10</sup> and either the Arg to Glu mutation at position 180 or, in addition, the Arg to Gly mutation at position 269. The resulting plasmids were cut with *EcoRV* within LEU2 and transformed into the diploid sectoring-assay strain AVY125 (*MATa/MATα SCC1/sccl::Kan<sup>R</sup> ura3-52/ura3-52 ade2-101/ade2-101 TRP1/trp1-Δ1 leu2-Δ1/ leu2-Δ1 CFVII (RAD2.d) URA3 SUP11*)<sup>19</sup> lacking one of two copies of SCC1. Haploid derivatives of the resulting transformants, verified for the presence of the intended SCC1 alleles, were the strains AVY129 (*MATa sccl::Kan<sup>R</sup> leu2::SCC1<sup>R180E</sup>::LEU2 ura3-52 ade2-101 trp1-Δ1 CFVII (RAD2.d) URA3 SUP11*) and AVY130 (*MATa sccl::Kan<sup>R</sup> leu2::SCC1<sup>R180E R269G</sup>::LEU2 ura3-52 ade2-101 trp1-Δ1 CFVII (RAD2.d) URA3 SUP11*).

## Experiments with dipeptides

We grew *S. cerevisiae* K6843 (*UBR1*)<sup>8</sup> and AVY115 (*ubr1Δ*) at 30 °C in YP (yeast-peptone) medium containing 2% raffinose as the carbon source to an absorbance at 600 nm (*A*<sub>600</sub>) ≈ 0.5. Cells were arrested at metaphase by nocodazole (Sigma), at 15 μg ml<sup>-1</sup>. The P<sub>GAL1</sub>-10 promoter was induced by the addition of galactose to 2% for 30 min. We added specific dipeptides (Sigma) to 5 mM, and incubated 2-ml cultures for another 30 min, followed by the addition of an equal volume of 0.3 M NaCl, 0.1 M NaF, 20 mM EDTA, 2 mM Na<sub>2</sub>S<sub>2</sub>O<sub>8</sub> (pH 8.2). Cells were pelleted by centrifugation, resuspended in 0.2 ml of the lysis buffer (10% glycerol, 50 mM NaOH, 2% SDS, 5% β-mercaptoethanol), and heated at 100 °C for 5 min. The suspension was titrated with 1 M HCl to pH 7, and centrifuged at 12,000g for 2 min. Samples of the supernatants (10 μl) were subjected to SDS-PAGE (10% gel), followed by immunoblotting with a 1:1000 dilution of anti-Myc 9E10 antibody (Berkeley Antibody Co.) The blot was incubated with a 1:2000 dilution of the goat anti-mouse horseradish peroxidase (HRP) conjugate, and was developed using ECL reagents (Amersham).

## Pulse-chase analysis

*S. cerevisiae* JD52 and JD55, carrying plasmids that expressed <sup>f</sup>DHFR-Ub-Arg-SCC1<sup>269-566f</sup> or <sup>f</sup>DHFR-Ub-Met-SCC1<sup>269-566f</sup> from the P<sub>GAL1</sub> promoter, were grown at 30 °C to *A*<sub>600</sub> ≈ 1 in SR(-ura) medium with auxotrophic supplements and 2% raffinose as the carbon source. P<sub>GAL1</sub> was induced by adding galactose, to 2% (SRG(-ura) medium), for 1 h. Cells from a 20-ml culture were gathered by centrifugation, washed with 0.8 ml SRG(-ura), resuspended in 0.4 ml of SRG(-ura), and labelled for 3 min with 0.16 mCi of <sup>35</sup>S-Express (NEN), followed by centrifugation and resuspension of cells in SD(-ura) containing 4 mM L-methionine and 2 mM L-cysteine. We took 0.1 ml samples at the indicated time points and processed them for immunoprecipitation with anti-Flag M2 agarose beads (Sigma), followed by SDS-PAGE, as described<sup>25</sup>.

## Experiments with growth inhibitors

We grew *S. cerevisiae* AVY111, which expresses Met-SCC1<sup>269-566f</sup> from the P<sub>GAL1</sub> promoter, and the congenic parental strain AVY110 at 30 °C in YP medium containing 2% raffinose as the carbon source to *A*<sub>600</sub> ≈ 0.5. To separate samples of the two cultures, were added α-factor (to 1 μg ml<sup>-1</sup>), hydroxyurea (to 0.1 M) or nocodazole (to 15 μg ml<sup>-1</sup>), followed by incubation for 2.5 h. One half of each of the resulting samples was then incubated further in the same medium for 1.5 h; the other half had galactose added to 2%, followed by a 1.5-h incubation. Cells were gathered by centrifugation, washed with water, and plated onto dextrose-containing synthetic-media (SD) plates.

## Co-immunoprecipitation/immunoblotting assay

JD52 cells carrying two vectors (pRS415 and pRS416<sup>25</sup>), or pRS415 and p416UPRSCC1M (expressing Flag-tagged Met-SCC1<sup>269-566f</sup>), or pRS416 and p415GalSMC1 (expressing RGS/His<sub>6</sub>-tagged SMC1), or p415GalSMC1 and p416UPRSCC1M together, were grown in the raffinose-containing SR medium to *A*<sub>600</sub> ≈ 1. Galactose was added to 2%, followed by incubation for 2 h, preparation of extracts<sup>25</sup>, immunoprecipitation with anti-Flag M2 beads, SDS-PAGE (10% gel), and immunoblotting, separately, with anti-Flag and anti-RGS/His<sub>6</sub> antibodies (Qiagen).

## Sectoring assay

*S. cerevisiae* YPH277 contained the defective *ade2-101* ochre allele of *ADE2*, which was suppressed by the SUP11 suppressor transfer RNA gene, carried on a non-essential, centromere-containing, ~90-kilobase chromosome fragment<sup>19</sup>. YPH277 cells that retained the SUP11-containing chromosome formed white colonies; the loss of this chromosome resulted in a red sector or sectors. The UBR1 gene was deleted in the YPH277 background, and the resulting *ubr1Δ* strain YPH277HR1 was compared with YPH277. Cells were grown in the uracil-lacking SD medium that retained the SUP11-containing chromosome, followed by plating on YPD plates.

Received 20 November 2000; accepted 17 January 2001.

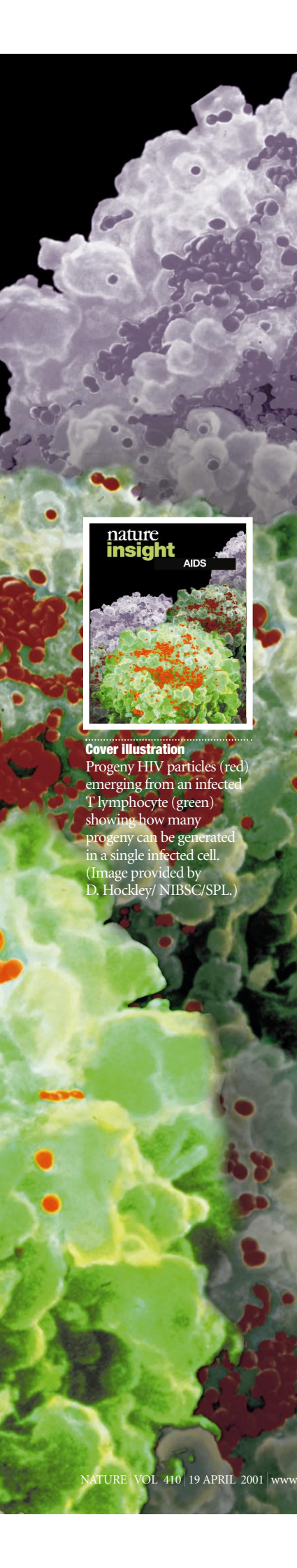
- Nasmyth, K., Peters, J. M. & Uhlmann, F. Splitting the chromosome: cutting the ties that bind sister chromatids. *Science* **288**, 1379–1384 (2000).
- Koshland, D. & Guacci, V. Sister chromatid cohesion: the beginning of a long and beautiful relationship. *Curr. Opin. Cell Biol.* **12**, 297–301 (2000).
- Yanagida, M. Cell cycle mechanisms of sister chromatid separation: roles of Cut1/separin and Cut2/securin. *Genes Cells* **5**, 1–8 (2000).
- Hirano, T. Chromosome cohesion, condensation, and separation. *Annu. Rev. Biochem.* **69**, 115–144 (2000).
- Dej, K. J. & Orr-Weaver, T. L. Separation anxiety at the centromere. *Trends Cell Biol.* **10**, 392–399 (2000).
- Pidoux, A. L. & Allshire, R. C. Centromeres: getting a grip of chromosomes. *Curr. Opin. Cell Biol.* **12**, 308–319 (2000).
- Biggins, S. & Murray, A. W. Sister chromatid cohesion in mitosis. *Curr. Opin. Genet. Dev.* **9**, 230–236 (1999).
- Uhlmann, F., Lottspeich, F. & Nasmyth, K. Sister-chromatid separation at anaphase onset is promoted by cleavage of the cohesin subunit Scc1. *Nature* **400**, 37–42 (1999).
- Varshavsky, A. The N-end rule: functions, mysteries, uses. *Proc. Natl Acad. Sci. USA* **93**, 12142–12149 (1996).
- Uhlmann, F., Wernic, D., Poupard, M.-A., Koonin, E. V. & Nasmyth, K. Cleavage of cohesin by the CD-clan protease separin triggers anaphase in yeast. *Cell* **103**, 375–386 (2000).
- Varshavsky, A. The ubiquitin system. *Trends Biochem. Sci.* **22**, 383–387 (1997).
- Hershko, A. & Ciechanover, A. The ubiquitin system. *Annu. Rev. Biochem.* **76**, 425–479 (1998).
- Bachmair, A., Finley, D. & Varshavsky, A. *In vivo* half-life of a protein is a function of its amino-terminal residue. *Science* **234**, 179–186 (1986).
- Suzuki, T. & Varshavsky, A. Degradation signals in the lysine-asparagine sequence space. *EMBO J.* **18**, 6017–6026 (1999).
- Xie, Y. & Varshavsky, A. The E2-E3 interaction in the N-end rule pathway: the RING-H2 finger of E3 is required for the synthesis of multiubiquitin chain. *EMBO J.* **18**, 6832–6844 (1999).
- Turner, G. C., Du, F. & Varshavsky, A. Peptides accelerate their uptake by activating a ubiquitin-dependent proteolytic pathway. *Nature* **405**, 579–583 (2000).
- Varshavsky, A. Ubiquitin fusion technique and its descendants. *Methods Enzymol.* **327**, 578–593 (2000).
- Turner, G. C. & Varshavsky, A. Detecting and measuring cotranslational protein degradation *in vivo*. *Science* **289**, 2117–2120 (2000).
- Spencer, E., Gerring, S. L., Connelly, C. & Hieter, P. Mitotic chromosome transmission fidelity mutants in *Saccharomyces cerevisiae*. *Genetics* **124**, 237–249 (1990).
- Buonomo, S. B. C. *et al.* Disjunction of homologous chromosomes in meiosis I depends on proteolytic cleavage of the meiotic cohesin Rec8 by separin. *Cell* **103**, 387–398 (2000).
- Tomonaga, T. *et al.* Characterization of fission yeast cohesin: essential anaphase proteolysis of Rad21 phosphorylated in the S phase. *Genes Dev.* **14**, 2757–2770 (2000).
- Mulder, L. C. F. & Muesing, M. A. Degradation of HIV-1 integrase by the N-end rule pathway. *J. Biol. Chem.* **275**, 29749–29753 (2000).
- Sijts, A. J., Philip, I. & Pamer, E. G. The *Listeria monocytogenes*-secreted p60 protein is an N-end rule substrate in the cytosol of infected cells. Implications for major histocompatibility complex class I antigen processing of bacterial proteins. *J. Biol. Chem.* **272**, 19261–19268 (1997).
- Waizenegger, I. C., Hauf, S., Meinke, A. & Peters, J.-M. Two distinct pathways remove mammalian cohesin from chromosome arms in prophase and from centromeres in anaphase. *Cell* **103**, 399–410 (2000).
- Ghislain, M., Dohmen, R. J., Levy, F. & Varshavsky, A. Cdc48p interacts with Ufd3p, a WD repeat protein required for ubiquitin-mediated proteolysis in *Saccharomyces cerevisiae*. *EMBO J.* **15**, 4884–4899 (1996).
- Mumberg, D., Muller, R. & Funk, M. Regulatable promoters of *Saccharomyces cerevisiae*—comparison of transcriptional activity and their use for heterologous expression. *Nucleic Acids Res.* **22**, 5767–5768 (1994).

## Acknowledgements

We are grateful to G. Turner, H.-R. Wang, F. Du, V. Ellison, A. Murray, D. Morgan, B. Stillman and P. Hieter for strains and plasmids. We thank M. Budd, W. Shuo and members of the Varshavsky laboratory, particularly F. Du and G. Turner, for helpful discussions. This work was supported by grants to A.V. from the NIH. H.R. is a Fellow of the Leukemia and Lymphoma Society. K.N. was supported by the Austrian Industrial Research Promotion Fund.

Correspondence and requests for materials should be addressed to A.V. (e-mail: avarsh@caltech.edu).





# nature insight

## AIDS

**T**he scale of the human immunodeficiency virus (HIV) epidemic is far greater than was predicted a decade ago. Already almost 20 million people have died of AIDS, over 30 million are currently living with HIV, and 16,000 new infections occur daily. It is clear that methods aimed at controlling the spread of the virus are failing to do so. Existing AIDS therapies are out of reach for most HIV-infected people in developing countries and, where available, they are limited by their toxicity and their cost. Furthermore, it is becoming increasingly apparent that current drugs are unable to purge the body of the virus and that discontinuation of triple therapy leads to a rapid rebound of plasma viraemia. Development of an effective HIV vaccine therefore remains one of the most pressing challenges facing modern medicine, and new types of drugs to treat already established disease are needed urgently.

In this month's *Nature Insight* we examine the biology of the virus and the disease it causes, give an overview on the current status of the AIDS pandemic, and review the efforts underway to control and curb it.

Robin Weiss introduces HIV on page 963 and discusses how the continuing epidemic could lead to a general increase in opportunistic infections. On page 968, Peter Piot and colleagues review the current AIDS global pandemic, covering the incidence and patterns of infection, the impact of disease on the socioeconomic make-up of affected countries, and potential strategies to control the rate of infection. On page 974, Mike McCune offers a comprehensive overview on the underlying causes for the progressive and relentless depletion of CD4<sup>+</sup> T cells, summarizing competing viewpoints and proposing a model that might resolve the current debate of 'accelerated destruction' versus 'regenerative failure'. Andrew McMichael and Sarah Rowland-Jones continue on page 980 with a review on the mechanisms that account for the fact that cytotoxic T cells — although effective in short- or medium-term containment of the virus — are ultimately unable to control infection. On page 988, Stuart Lipton and colleagues review molecular pathways that give rise to neuronal injury and apoptosis in HIV-associated dementia. Douglas Richman reviews HIV chemotherapy on page 995. Although largely out of reach to the developed world, but new anti-HIV compounds are needed urgently to combat emerging viral resistance and reduce the side effects associated with the toxicities of currently available drugs. Gary Nabel ends this collection of review articles with an overview on page 1002 on the scientific challenges facing HIV vaccine development and how those challenges are being met by a growing commitment of the research community.

We are pleased to acknowledge the financial support of Bristol-Myers Squibb in producing this Insight. As always, *Nature* carries the sole responsibility for all editorial content and peer-review. We hope that both general readers as well as experts in the field will find these articles useful and informative.

**Ursula Weiss** Senior Editor

**Chris Surridge** Insight Programme Editor  
**Liz Allen** Publisher

# Gulliver's travels in HIVland

Robin A. Weiss

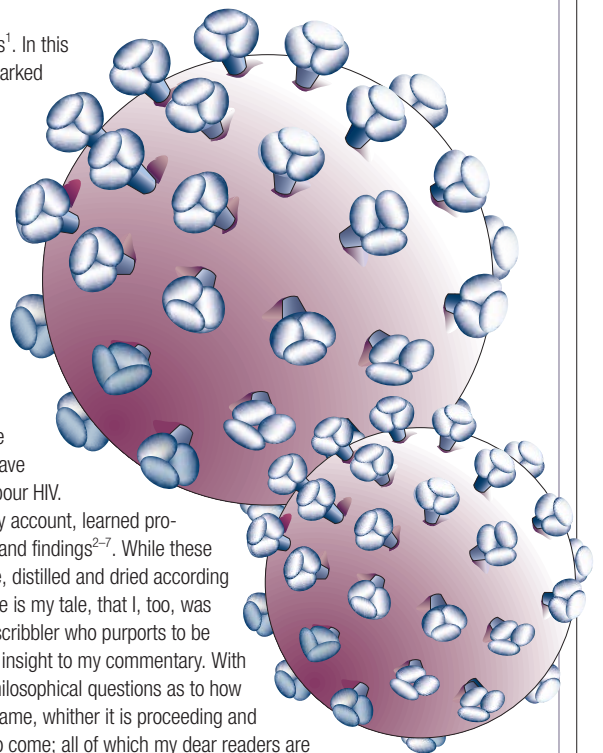
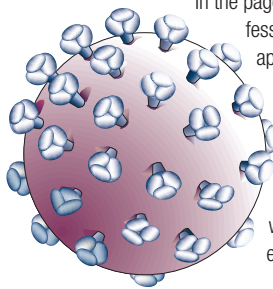
Wohl Virion Centre, Department of Immunology & Molecular Pathology, University College London, Windeyer Institute of Medical Sciences, 46 Cleveland Street, London W1T 4JF, UK (e-mail: r.weiss@ucl.ac.uk)

**The emergence of HIV and AIDS is narrated here through the eyes of the legendary Irish traveller Gulliver, observing the replication, cross-species origin, evolution, diversity and transmission of HIV. Ethical problems of vaccine trials, the social impact of AIDS, and prospects for its prevention, including the development of topical virucidal lotions, are discussed. The existence of a growing proportion of HIV-infected, immunocompromised children and adults may significantly affect current immunization programmes and the evolution of opportunistic infections.**

I, Lemuel Gulliver, have observed wondrous phenomena in many lands<sup>1</sup>. In this my latest account, I shall endeavour to convince you that we are embarked upon a doleful new adventure that is only now beginning to unfold<sup>2</sup>. My story concerns a creature even smaller than the Lilliputians, indeed so minute as to be invisible, named — after quarrelsome debate among a band of pundits in 1986 — the human immunodeficiency virus (HIV). This virus invades the body and slowly corrupts its defences<sup>3,4</sup>, so that without intervention by apothecary's powders<sup>5</sup>, or a modification of Mr Jenner's vaccine<sup>4,6</sup>, the infected subject will eventually perish, on account of invasion by other microbes, and a general wasting of the body, or the mind<sup>7</sup>, a collation of afflictions called AIDS<sup>8</sup>.

According to the esteemed comptrollers of estimates in Geneva<sup>2</sup>, during the first year of the third millennium, this fourth horseman of the apocalypse has caused the demise of some six million people. Since the pestilence first came to notice exactly 20 years ago<sup>9</sup>, some 23 million have been slain by it, and 37 million women, men and children currently harbour HIV.

In the pages of this journal, following hard on my account, learned professors of HIVland survey the latest facts and findings<sup>2-7</sup>. While these appendices contain scientific knowledge, distilled and dried according to usual academic practice, so strange is my tale, that I, too, was implored by my editor to engage a scribbler who purports to be an expert on these matters, to add insight to my commentary. With his help, I shall confine myself to philosophical questions as to how HIV procreates in cells, whence HIV came, whither it is proceeding and what HIVland may resemble in years to come; all of which my dear readers are enjoined to view as idle speculation, unsure prediction or wild conjecture.



## How HIV commandeers host cells

Like all viruses, HIV is a parasite that replicates within living cells of the host (Fig. 1). HIV has nine genes and belongs to the lentivirus genus of retroviruses. It carries two RNA copies of its genome within virus particles. Viral reverse transcriptase converts them to one DNA copy within the infected cell (a form of meiosis involving genome recombination), which enables HIV to be integrated into the host DNA and to use the cell's genetic machinery to make new virus. Reverse transcriptase was the first target of antiviral drugs in clinical use<sup>5</sup>, the second being the protease enzyme that cleaves precursor proteins into components of the viral core during particle assembly. Other promising targets for anti-HIV drug development include the interaction with cell-surface receptors and integration of the viral DNA into host chromosomal DNA.

HIV infects cells mainly of the immune system<sup>8</sup>. T-helper

lymphocytes express the CD4 antigen to which HIV attaches. Chemokine receptors serve as co-receptors that guide the viral envelope glycoproteins into a conformation permitting membrane fusion and entry into the cell. Although many co-receptors have been defined in culture, only CCR5 and CXCR4 seem to be used *in vivo*, by HIV strains having an 'R5' or an 'X4' tropism, respectively<sup>10</sup>. Besides CD4<sup>+</sup> lymphocytes, macrophages and dendritic cells also harbour HIV. Macrophages are an important reservoir of infection, including the microglia in the brain<sup>7</sup>. Dendritic cells bind HIV through DC-SIGN receptors and carry it from mucosal ports of entry to the lymph nodes where they activate lymphocyte infection<sup>11</sup>. High levels of virus replication and cell destruction and turnover occur at all stages of infection<sup>3,5</sup>, although it takes years before the CD4<sup>+</sup> cell count falls below the threshold for symptomatic immunodeficiency to become manifest.

Once upon a time, in the early twentieth century, there was a virus — the simian immunodeficiency virus or SIV<sub>CPZ</sub> — that lived in harmony with its host, the chimpanzee. This host resembled the yahoo, described in my voyage to the houghnhms<sup>1</sup>, save it was hairier, and of a milder disposition. One night, SIV<sub>CPZ</sub> had a premonition that over the course of the next 100 years its gentle host would so diminish in numbers as to be in grave danger of extinction. Being wholly reliant on the chimpanzee for its own procreation, SIV<sub>CPZ</sub> would not survive unless it found a new home.

Now, there was a legend so fanciful one durst not repeat it for fear of ridicule; yet it be so strong in folk memory that one plucks the courage to declare it. Namely, a distant cousin to SIV, yellow fever virus (YFV), had boarded a six-legged, two-winged little monster called *Aedes* that supped on the blood of the hairy ape, then drank from the yahoo, thus transmitting YFV. These newly infected naked apes were captured by cruel kinsmen from the north and were bound and shackled to sailing vessels that crossed the great ocean westerly for 40 days and nights, to disembark upon a new shore. Ever-helpful *Aedes* assisted YFV to escape and colonize numerous New World monkeys and yahoos, and thrive among them. Each summer, YFV made forays to the north, slaying thousands in Memphis, Philadelphia and even New Amsterdam until finally, 101 years ago, Colonel Walter Reed and Dr Carlos Finlay demonstrated the role of *Aedes* and brought the advancing enemy to a halt.

To its alarm, SIV<sub>CPZ</sub> found it was unable to emulate YFV because *Aedes*, and her sisters *Culex* and *Anopheles*, declined to transmit it. No one is entirely sure how the virus managed to cross into its human host to become HIV-1. Indeed, its possible routes of cross-species transmission have recently been keenly disputed at the Academy of Lagado<sup>12</sup> (described in my voyage to Laputa<sup>1</sup>) and contamination of oral polio vaccine no longer seems plausible<sup>13,14</sup>. But cross the species 'barrier' it did. Likewise, SIV<sub>SM</sub> of sooty mangabey monkeys found its way to humans as HIV-2.

## Adaptation and spread

The cross-species transfer of HIV-1 and HIV-2 from chimpanzees and mangabeys into humans resulted in a change of virulence, as the natural simian hosts do not develop AIDS. Yet the pathogenesis of primate immunodeficiency viruses seem oddly uncoupled from their transmission dynamics. For example, asymptomatic sooty mangabeys carry viral loads as high as the macaques and humans that succumb to AIDS from infection by essentially the same virus<sup>15</sup>. This change of virulence may therefore be determined more by the host response to infection than by properties of the virus. One possible explanation is that HIV infection activates the human immune system, providing a larger pool of cells permissive for ongoing infection and for apoptosis (programmed cell death)<sup>16</sup>. The eventual depletion of CD4<sup>+</sup> T-lymphocytes<sup>3</sup> finally tips the balance of power between virus and the immune system in favour of the virus. If we understood better how the host controls SIV infection in mangabeys and chimpanzees, we might find ways to effect HIV control in humans.

HIV-2 seems to be less pathogenic than HIV-1 and to cause disease more slowly<sup>17</sup>. But this observation may mask a bimodal spectrum of virulence in which some individuals carrying HIV-2 progress to AIDS at a similar rate to those with HIV-1, whereas a higher proportion are long-term non-progressors. It is also noteworthy that among the HIV-2-infected people who do develop AIDS, brain disease is more common<sup>18</sup>.

HIV-1 and -2 probably encountered many hurdles in adapting to their new host. Some expeditions to humankind may have petered out altogether; others like HIV-1 group N and O remained geographically close to the point of cross-species transfer. But HIV-1 group M was more adventurous, taking every opportunity to transmit its progeny by exploring the highways and intimate byways of its host's behaviour — crossing sexually from man to woman, woman to man

and man to man, crossing vertically from mother to child, and horizontally through hollow needles.

## HIV diversity

Needles especially may have unwittingly aided the early dispersion of HIV, owing to the widespread re-using of needles and syringes in Africa in the mid-twentieth century<sup>19</sup>. Once HIV became established in its new host, overland trucks, overseas troops and airlines enabled a far more rapid and widespread dispersion than mosquitoes could have effected. HIV-1 group M diversified into distinct subtypes or clades, A–H, with subtype B colonizing the Americas, subtype C moving south to the Cape and north to the Horn of Africa, and subtype E east to Thailand. This outward radiation of HIV-1 group M genomes can be plotted as a starburst<sup>20</sup>, but when the clades crossed paths, HIV-1 recombined, so that some genomes have a highly complex genealogy<sup>21,22</sup>. Even recombinants between groups M and O have been recorded<sup>23,24</sup>.

My fable related how the precursor to HIV-1 escaped extinction in an ever dwindling host population, currently estimated to be less than 150,000 chimpanzees, separated into isolated troupes. HIV can now sample a new horde of almost six billion, among whom it currently resides in 37 million, or approximately 1 in 162 humans. And by weakening the host's immune system, HIV opened the door to numerous other microbes — the opportunistic infections by viruses, bacteria, fungi and animal parasites.

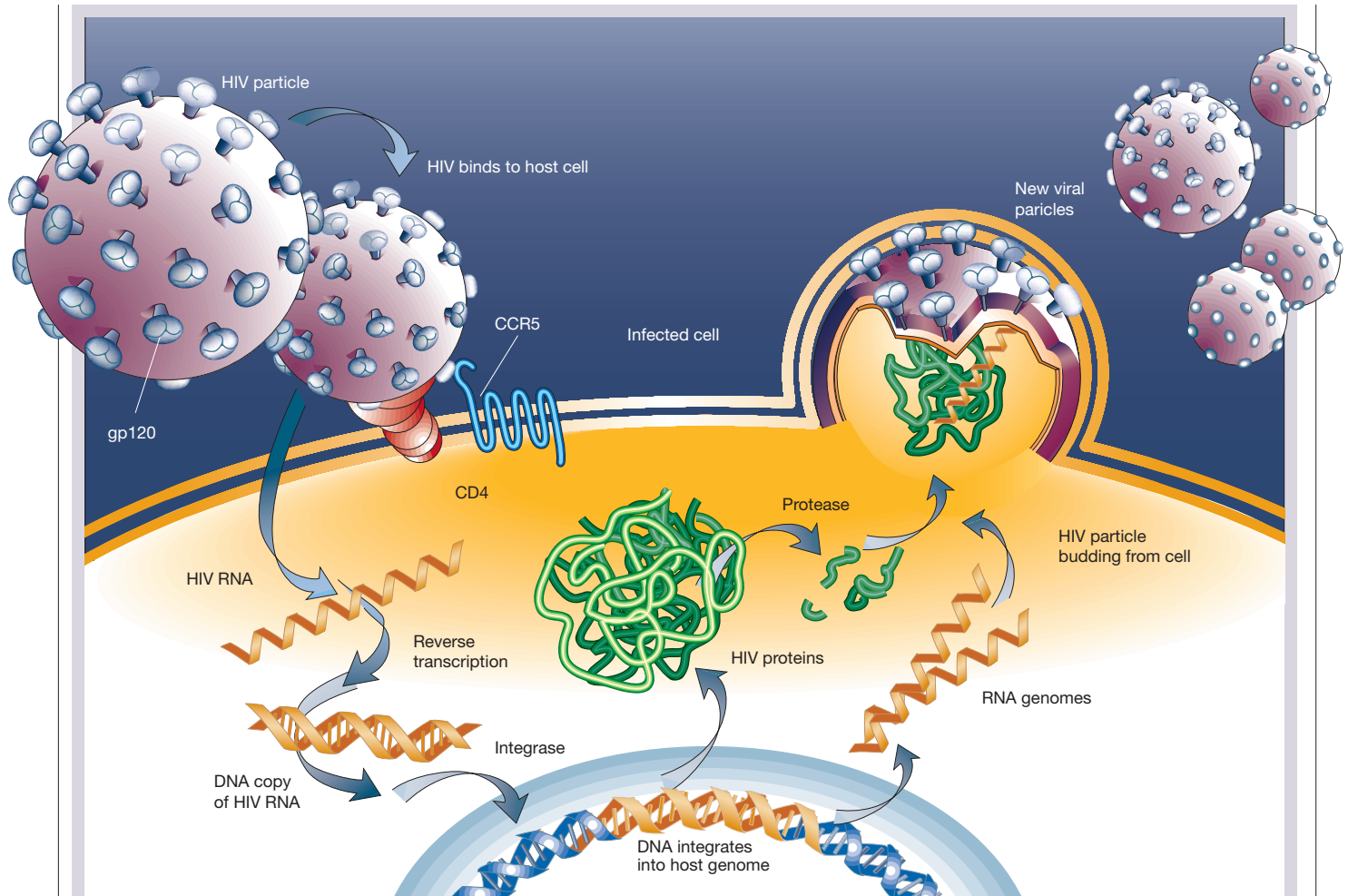
Now, I shall surely be admonished by some high authority if I would have you believe that a simple virus were capable of premeditated purpose, which has no place in HIV's affairs. Rather HIV has followed the precepts postulated by Charles Darwin and Alfred Russel Wallace, whereby the fittest shall survive and multiply, passing on their fitness traits. Neither is HIV capable of altruistic behaviour towards unrelated microbes, according to the precepts of Bill Hamilton<sup>25</sup>. So the term 'opportunistic infection' aptly describes the exploitation of the expanding ecological niche provided by HIV.

Is HIV evolving in the sense of changing its phenotype, mode of replication, virulence and transmission<sup>22,26</sup>? Or, is it merely expanding, with genetic diversity representing more noise than biological signal? The answer is both, for the basic molecular biology of replication remains constant (Fig.1) while the explosion of HIV into the new terrain of the human body and the human population has allowed an unprecedented degree of genetic diversity, upon which natural selection can then play. HIV selection occurs rapidly in response to antiretroviral drugs<sup>5</sup> and immune attack<sup>4</sup>, compounding the problem of vaccine development<sup>6</sup>. It occurs when the virus adapts to use new co-receptors for entry to the host cell, but an identical mutation can both alter cell tropism and lead to escape from neutralization<sup>27</sup>. So it is a moot point as to which is exerting the selective pressure.

There is much we do not yet understand about HIV variation. Is there selection of variants for different compartments such as the lymph nodes and brain, or do the differences reflect stochastic founder effects? Do the envelope variants that emerge in late-stage infection, with a tendency to switch from R5 to X4 phenotype, really exacerbate progression to AIDS, or are they opportunistic variants that cannot emerge until CD4 T-cell counts fall below a threshold level? Why have X4 viruses not emerged in subtype C<sup>28</sup>? Why is the R5 phenotype usually reset at transmission? One explanation is that only R5 viruses can bind to or infect dendritic cells such as Langerhans cells in the mucosa<sup>11</sup>, yet R5 viruses are also selected following par-enteral transmission. Are some subtypes more virulent than others? Could HIV-1 eventually become attenuated<sup>22</sup>?

HIV is an ideal organism for functional genomics modelling; HIV is a ready-made experiment in maximal mutagenesis while maintaining function and replicative fitness. This has already been exemplified for HIV-1 reverse transcriptase and protease, for which





**Figure 1** The human immunodeficiency virus. Replication cycle of HIV showing multiple steps available for attack by antiretroviral drugs.

there exist crystal structures, resistant variants to numerous drugs, a vast sequence database and relative fitness data. The degree of genomic diversity that HIV generates in a single infected individual can be greater than the worldwide diversity of influenza A virus during an epidemic<sup>29</sup>. Where, then, can HIV take us, with millions of infected people each possessing such viral diversity?

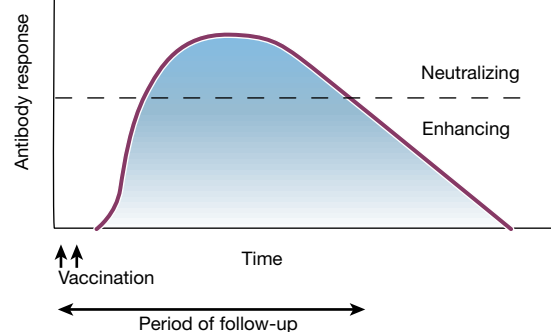
How conservative is the HIV phenotype and what new tricks might HIV learn? One nightmare scenario for us would be if HIV were to change its mode of transmission. If *Yersinia pestis* can switch from flea-borne bubonic plague to the air-borne pulmonary form of the disease, could HIV also sample new transmission dynamics — adding saliva, aerosol or arthropod vectors to the sex and blood it already enjoys? Forget the mosquitoes beloved of yellow fever, but consider ticks and biting bugs. Can we be sure that it is beyond the ingenuity of HIV to travel aboard the mouthparts (analogous to dirty needles) during interrupted feeding of common bugs such as the cone-nosed *Rhodnius* in Brazil, or the bed-bug *Cimex* in Russia? After all, horseflies or clegs transmit the lentivirus, causing equine infectious anaemia<sup>30</sup>. Identifying a new route of transmission may be overlooked unless surveillance methods are designed to detect it. Sentinel cases might be children with HIV-negative parents.

### Prevention and intervention

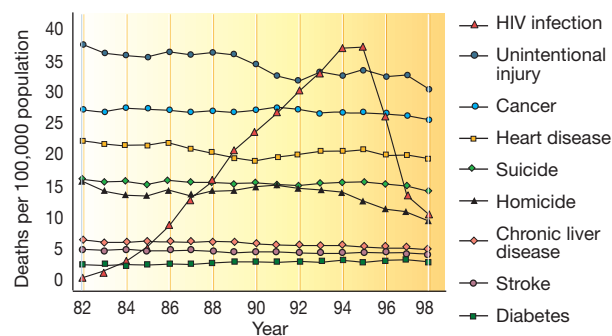
Uneven as the landscape is, HIVland looks set to extend its borders to grasp tens of millions more infected people<sup>31</sup>. There are, however, some means of slowing its expansion, and some badly affected communities have managed to reduce the rate of HIV transmission through health education<sup>2</sup>. Restraint on the number of sexual

partners, treatment of common sexually transmitted infections (which increase susceptibility to HIV) with antibiotics and antivirals, promoting condom use, and clean needle-exchange programmes all help to curb HIV transmission.

The identification of protective factors of HIV transmission does not necessarily make prevention easy. For instance, it is now clear that circumcised men in Africa have a significantly lower probability of acquiring HIV infection<sup>32</sup>. These epidemiological findings make



**Figure 2** Scheme showing how a humoral immune response to gp120 vaccine may switch from providing protection to enhancing the risk of infection, perhaps after the surveillance period has ended.



**Figure 3** Trends in annual rates of leading causes of death among adults aged 25–44 years in USA over the period 1982–1998. The data from 1998 are preliminary. (Source: Centers for Disease Control and Prevention.)

biological sense, given the high density of Langerhans cells in the prepuce<sup>33</sup>. Although some communities that did not previously practise circumcision are beginning to do so, promulgating such an intervention as public-health policy is fraught with problems. Not only is circumcision a sensitive cultural issue, but a false sense of security might be counterproductive. Although circumcision reduces the chances of HIV infection in men by approximately 60%, it might lead to a rise in HIV transmission through higher-risk sex if men considered themselves to be completely protected. Condoms would be more effective if only men everywhere could be persuaded to adopt them, and to adhere to their use regularly<sup>34</sup>.

The Cinderella of HIV research is the development of topical microbicides or virucides. Compared with the research funding available to her big sisters, antiretroviral drugs and vaccines, virucides need to attract more support for research and development<sup>34</sup>. If virucides can be incorporated successfully into vaginal gels, foams and suppositories, they would enable women to provide a means to improving sexual health for both sexes. Whereas the spermicide nonoxonyl-9 is too inflammatory, certain sulphated polysaccharides show promise for formulation as anti-HIV molecules for topical application<sup>35</sup>. Virucides could be our best hope for preventing HIV infection until we can deliver a safe, efficacious, affordable vaccine.

Although HIV elicits strong immune responses, these ultimately fail to control infection<sup>4</sup>. Nabel<sup>6</sup> catalogues the obstacles confronting an effective HIV vaccine, while pointing to some promising leads. Cohen<sup>36</sup> has documented the lack of coordination of HIV vaccine research and development. Even a vaccine with <50% efficacy could have a significant impact in reducing HIV transmission<sup>37</sup>. Yet a dilemma exists between purists who wish to see scientific proof of efficacy, and pragmatists who believe that the evidence will come only from conducting human trials<sup>36</sup>.

The pressure to conduct clinical trials of candidate vaccines is intense. An immunogen based on the gp120 envelope antigen has progressed to phase III clinical trials, after approval by international and local ethical committees, despite no evidence of efficacy, on the lilliputian logic that if it is available it must be used. Gp120 stimulates a humoral response, including weak neutralization of the HIV strain from which it is derived. This type of vaccine might conceivably be worse than useless, as it might elicit antibodies that enhance infection by targeting macrophages and dendritic cells via complement and Fc receptors<sup>8</sup>. The trial design is unlikely to reveal any increased risk because the enhancement may occur after the period of follow-up among high-risk vaccinees has ended (Fig. 2); and by targeting people illicitly injecting drugs there is likely to be significant loss to follow-up by those at greatest risk. When this writer ventured to raise these points at an AIDS vaccine conference last May, he was politely but firmly rebuked by an envoy from one of the worst afflicted countries in Asia on the grounds that 'something must be done' for his people, a

view that won overwhelming applause from the participants. One remains concerned that the heart overrules the head, although we earnestly hope that these fears will prove to be unfounded.

To most readers of *Nature*, AIDS research has followed a steady progress of science successfully translated into medical practice. AIDS was first noted<sup>9</sup> in 1981. Within two years HIV was identified, and by late 1985, serological screening was in place to prevent further infection through blood transfusion and blood products. By 1986, the first antiretroviral drug, zidovudine, was in clinical trial, culminating in the success of combination chemotherapy<sup>5</sup> based on rational drug design, which has led to a 60% reduction of AIDS mortality in the United States (Fig. 3). Our knowledge of HIV and of cell biology is immense, although vaccine development has been much slower than forecast.

Among the people most affected by HIV, Western gay men seized the opportunity to convert stigma and fear into empowerment to set the agenda for research and public health, often knowing more about their condition than their doctors. But for the 90% of HIV-infected people in developing countries there has been little or no access to these scientific advances. There is currently a North–South dispute between protecting the patents of the pharmaceutical companies and demanding access to cheaper generic drugs as a human right. But even generic drugs are beyond the health budgets of the worst affected nations. Besides, if 20% of the US population were HIV-infected, its health management and insurance systems would collapse.

### Culture and belief in HIVland

Sometimes the drugs are shunned even when offered at no cost. After HIV swept across southern Africa in the 1990s, some leaders blamed poverty rather than HIV for AIDS, finding scapegoats to attack, just as when the Black Death came to Europe in 1347<sup>38,39</sup>. One leant his ear to siren voices singing that HIV was harmless and that AIDS was not transmissible, so encouraging the misconception that AIDS in Africa is a distinct disease, contrary to all the evidence<sup>40,41</sup>.

HIV will not drive humans to extinction, for even rabbits survived the 99% mortality of myxomatosis. But HIV has already reduced life expectancy<sup>2</sup> and will destabilize society by selectively removing its young adults.

Who, therefore, can foretell what changes in belief<sup>42</sup> we may witness in the coming years, as the nature and extent of HIV's destruction finally becomes apparent. Did not the sixth-century Justinian plague presage the rise in ascetic religions — Christianity in Europe, Islam in the Near East and Buddhism in the Far East? Did not the Pope in Avignon first encourage, then try to suppress, the Flagellants<sup>39</sup> who rose up in 1348 to counter the Black Death that culled saints and sinners alike? Did not smallpox introduced by the Spanish 'centaurs' help to convince the demoralized Aztecs that Christ was more powerful than Quetzalcóatl? Did not the pandemic of greatpox<sup>43</sup> (syphilis) that cut a swathe across Europe and Asia from 1495–1530 influence the Reformation? Did its devastation not help to persuade the Tokugawa Shogunate in Japan to quarantine their islands from the outside world for nigh on 300 years? Although Bynum<sup>44</sup> has cautioned *Nature's* readers not to swallow too facile an interpretation of past pestilence in the light of AIDS today, it would be surprising if the HIV pandemic did not effect major changes in our mores. So will the band play on<sup>45</sup>, or will the pendulum swing away from our contemporary society, as the religious right press AIDS to their cause of just deserts, and 'ecofascists' bless HIV for curbing human population growth?

### Effect of HIV on infectious disease

The impact of HIV/AIDS on other infectious diseases will be enormous. First, opportunistic infections, which seldom cause serious disease in immunocompetent people, are frequently the sentinels of

AIDS<sup>9</sup>. Over 100 opportunistic infections by viruses, bacteria, fungi and protozoa have been associated with AIDS. Disease by known pathogens such as tuberculosis also becomes exacerbated in AIDS. Antibiotic resistance to bacteria is more frequently found, and HIV-infected adults may regain susceptibility to childhood infections such as *Streptococcus pneumoniae*<sup>46</sup>. Thus the HIV-infected population can spread virulent strains of pathogens such as tuberculosis to others who are not necessarily at risk of HIV itself.

Second, HIV infection affects public-health immunization programmes. For instance, multivalent pneumococcal vaccine is ineffective in HIV-infected Africans<sup>47</sup>. Live 'attenuated' vaccines such as vaccinia, measles and oral polioviruses may become dangerous pathogens in the immunosuppressed person<sup>48–50</sup>. Moreover, HIV can convert acute infections into long-persistent ones<sup>50</sup>, as cell-mediated immune responses are crucial for clearance of virus infections like enteroviruses, measles, mumps and influenza. The World Health Organization needs to give further thought to the power of HIV to scupper its disease-eradication programmes<sup>49</sup>. With 1.5 million HIV-infected children in Africa, can we really eradicate polio within the next few years?

Third, AIDS could conceivably generate new types of epidemics. When the proportion of immunodeficient individuals in a population remains very low, the chance of direct transmission of opportunistic infections from one to another is even smaller. But HIV-infected people can act as incubators of microbes that previously relied on animal reservoirs: for example, the *Mycobacterium avium* complex, canine *Toxoplasma*, alphaviruses of birds, enteric infections of farm animals, and numerous other microbes not normally transmissible between humans.

With the AIDS pandemic, such microbes now have a new host population in which to play darwinian selection. Where 10% or more of a community are HIV-infected, direct transmission between immunodeficient individuals becomes plausible. Microbes that are poorly adapted for human infection could become well adjusted, first to the immunodeficient host and eventually to immunocompetent humans, provided they learn the tricks of human-to-human transmission. These could include free-living microbes from the environment<sup>51,52</sup>, including non-tuberculous *Mycobacteria*, *Legionella*, *Pseudomonas*, *Fusarium* and other fungi, as well as microbes and parasites from animal sources.

Free-living bacteria, fungi and protozoa now have 37 million and rising immunocompromised people in which to learn to become human parasites. The bacteria could acquire antibiotic resistance or adaptive, pathogenic genes from the plasmids, phages and mobile pathogenicity islands<sup>53</sup> of their human-adapted relatives, and yet escape the vaccines now in use. Moreover, animal infections could gain a firm foothold in the human population, particularly fast-evolving RNA viruses. The vast number of susceptible humans is a novel, unique window of opportunity for microbes originating from animals, the soil or water to evolve into human pathogens. This issue of the potential community impact of a massive immunodeficient reservoir therefore demands analysis by those who model infectious disease dynamics and evolution. Long after HIV itself is controlled (we hope) by immunization, new diseases may roam former HIVland. □

1. Swift, J. *Gulliver's Travels* (Penguin, London, 1967). [First published by B. Motte, London, 1726.]
2. Piot, P., Bartos, M., Ghys, P. D., Walker, N. & Schwartzländer, B. The global impact of HIV/AIDS. *Nature* **410**, 968–973 (2001).
3. McCune, J. M. The dynamics of CD4<sup>+</sup> T-cell depletion in HIV disease. *Nature* **410**, 974–979 (2001).
4. McMichael, A. J. & Rowland-Jones, S. L. Cellular immune responses to HIV. *Nature* **410**, 980–987 (2001).
5. Richman, D. D. HIV chemotherapy. *Nature* **410**, 995–1001 (2001).
6. Nabel, G. J. Challenges and opportunities for development of an AIDS vaccine. *Nature* **410**, 1002–1007 (2001).
7. Kaul, M., Garden, G. A. & Lipton, S. A. Pathways to neuronal injury and apoptosis in HIV-associated dementia. *Nature* **410**, 988–994 (2001).
8. Levy, J. A. *HIV and the Pathogenesis of AIDS* (American Society for Microbiology, Washington DC, 1998).
9. Centers for Disease Control & Prevention. Kaposi's sarcoma and *Pneumocystis pneumonia* among homosexual men in New York City and California. *Mort. Morb. Wkly. Rep.* **30**, 305–308 (1981).

10. O'Brien, S. J. & Moore, J. P. The effect of genetic variation in chemokines and their receptors on HIV transmission and progression to AIDS. *Immunol. Rev.* **177**, 99–111 (2000).
11. Geijtenbeek, T. B. et al. DC-SIGN, a dendritic cell-specific HIV-1-binding protein that enhances trans-infection of T cells. *Cell* **100**, 587–597 (2000).
12. Wain-Hobson, S. & Weiss, R. A. (eds) *The Origins of HIV and the AIDS Epidemic* (Phil. Trans. R. Soc. Lond. B **356**) (The Royal Society, London, 2001).
13. Blancou, P. et al. Polio vaccine samples not linked to AIDS. *Nature* (in the press).
14. Berry, N. et al. Analysis of oral polio vaccine CHAT stocks. *Nature* (in the press).
15. Kurth, R. & Norley, S. Why don't the natural hosts of SIV develop simian AIDS? *J. NIH Res.* **8**, 33–37 (1996).
16. Dalgleish, A. G. et al. The role of HIV gp120 in the disruption of the immune system. *Immunol. Lett.* **66**, 81–87 (1999).
17. Whittle, H. et al. HIV-2-infected patients survive longer than HIV-1-infected patients. *AIDS* **8**, 1617–1620 (1994).
18. Lucas, S. B. et al. The mortality and pathology of HIV infection in a west African city. *AIDS* **7**, 1569–1579 (1993).
19. Marx, P. A., Alcázar, P. G. & Drucker, E. Serial human passage of SIV unsterile injecting and the emergence of epidemic HIV in Africa. *Phil. Trans. R. Soc. Lond. B* **356** (in the press).
20. Korber, B. et al. Timing the ancestor of the HIV-1 pandemic strains. *Science* **288**, 1789–1796 (2000).
21. Vidal, N., Mulanga-Kabeya, C., Nzilambi, N., Delaporte, E. & Peeters, M. Identification of a complex env subtype E HIV type 1 virus from the Democratic Republic of Congo, recombinant with A, G, H, J, K, and unknown subtypes. *AIDS Res. Hum. Retroviruses* **16**, 2059–2064 (2000).
22. Malim, M. H. & Emerman, M. HIV-1 sequence variation, drift, shift, and attenuation. *Cell* **104**, 469–472 (2001).
23. Takehisa, J. et al. Human immunodeficiency virus type 1 intergroup (M/O) recombination in Cameroon. *J. Virol.* **73**, 6810–6820 (1999).
24. Peeters, M. et al. Characterization of a highly replicative intergroup M/O human immunodeficiency virus type 1 recombinant isolated from a Cameroonian patient. *J. Virol.* **73**, 7368–7675 (1999).
25. Hamilton, W. D. *Narrow Roads of Gene Land: Evolution of Social Behaviour* (Freeman, Oxford, 1995).
26. Sala, M. & Wain-Hobson, S. Are RNA viruses adapting or merely changing? *J. Mol. Evol.* **51**, 12–20 (2000).
27. McKnight, A. et al. Change in tropism upon immune escape by human immunodeficiency virus. *J. Virol.* **69**, 3167–3170 (1995).
28. Ping, L. H. et al. Characterization of V3 sequence heterogeneity in subtype C human immunodeficiency virus type 1 isolates from Malawi: underrepresentation of X4 variants. *J. Virol.* **73**, 6271–6281 (1999).
29. Korber, B. et al. in *The Changing face of HIV and AIDS* (eds Weiss, R. A., Adler, M. W. & Rowland-Jones, S. L.) (Br. Med. Bull., in the press).
30. Montelaro, R. C., Ball, J. M. & Rushlow, K. E. in *The Retroviridae* (ed. Levy, J. A.) 257–360 (Plenum, New York, 1993).
31. Carael, M., Buvé, A. & Awusabo-Asare, K. The making of HIV epidemics: what are the driving forces? *AIDS* **11**, S23–S31 (1997).
32. Weiss, H. A., Quigley, M. A. & Hayes, R. J. Male circumcision and risk of HIV infection in sub-Saharan Africa: a systematic review and meta-analysis. *AIDS* **14**, 2361–2370 (2000).
33. Szabo, R. & Short, R. V. How does male circumcision protect against HIV infection? *Br. Med. J.* **320**, 1592–1594 (2000).
34. Hill, R., Ryan, J., Stone, A. & Fransen, L. Vaginal microbicides for the prevention of HIV/AIDS: assessment of the potential market. *Int. J. Pharm.* **14**, 271–278 (2000).
35. McCormack, S., Hayes, R., Lacey, C. J. & Johnson, A. M. Microbicides in HIV prevention. *Br. Med. J.* **322**, 410–413 (2001).
36. Cohen, J. *Shots in the Dark: The Wayward Search for an HIV Vaccine* (Norton, New York, 2001).
37. Anderson, R. M. & Garnett, G. P. Low-efficacy HIV vaccines: potential for community-based intervention programmes. *Lancet* **348**, 1010–1013 (1996).
38. Lederberg, J. Infectious history. *Science* **288**, 287–293 (2000).
39. Ziegler, P. *The Black Death* (Penguin, London, 1970).
40. Durban Declaration. *Nature* **406**, 15–16 (2000).
41. Morgan, D. & Whitworth, J. A. The natural history of HIV-1 infection in Africa. *Nature Med.* **7**, 143–145 (2001).
42. Ranger, T. & Slack, P. (eds) *Epidemics and Ideas: The Historical Perception of Pestilence* (Cambridge Univ. Press, 1992).
43. Quétel, C. *History of Syphilis* (Blackwell, Oxford, 1990).
44. Bynum, W. F. Sexually transmitted retribution. *Nature* **407**, 943 (2000).
45. Shils, R. *And the Band Played On—Politics, People and the AIDS Epidemic* (St Martin's, New York, 1987).
46. Crewe-Brown, H. H. *Streptococcus pneumoniae* blood culture isolates from patients with and without HIV infection: alterations in penicillin susceptibilities and in serogroups or serotypes. *Clin. Inf. Dis.* **25**, 1165–1172 (1997).
47. French, N. et al. 23-Valent pneumococcal polysaccharide vaccine in HIV-1-infected Ugandan adults: double-blind, randomised and placebo controlled trial. *Lancet* **355**, 2106–2111 (2000).
48. Chitsike, I. & van Furth, R. Paralytic poliomyelitis associated with live oral poliomyelitis vaccine in a child with HIV infection in Zimbabwe. *Brit. Med. J.* **318**, 841–843 (1999).
49. Wood, D. J., Sutter, R. W. & Dowdle, W. R. Stopping poliovirus vaccination after eradication: issues and challenges. *Bull. World. Health Organ.* **78**, 347–357 (2000).
50. Moss, W. J., Cutts, F. & Griffin, D. E. Implications of the human immunodeficiency virus epidemic for control and eradication of measles. *Clin. Infect. Dis.* **29**, 106–112 (1999).
51. Shafer, R. W. & Sierra, M. F. *Mycobacterium xenopi*, *Mycobacterium fortuitum*, *Mycobacterium kansasii*, and other nontuberculous mycobacteria in an area of endemicity for AIDS. *Clin. Infect. Dis.* **15**, 161–162 (1992).
52. Mendelson, M. Fungal infections in the immunocompromised. *Microbiol. Today* **28**, 10–12 (2001).
53. Hacker, J. & Kaper, J. B. Pathogenicity islands and the evolution of microbes. *Ann. Rev. Microbiol.* **54**, 641–679 (2000).

#### Acknowledgements

The scribbler is indebted to V. Gant, P. Hale, A. McKnight, R. S. Tedder and S. Wain-Hobson for sceptical discussions. His AIDS research is supported by the Medical Research Council.



# The global impact of HIV/AIDS

Peter Piot, Michael Bartos, Peter D. Ghys, Neff Walker & Bernhard Schwartländer

Joint United Nations Programme on HIV/AIDS (UNAIDS), 20 avenue Appia, 1211 Geneva 27, Switzerland (e-mail: bartosm@unaids.org)

**The scale of the human immunodeficiency virus (HIV)/AIDS epidemic has exceeded all expectations since its identification 20 years ago. Globally, an estimated 36 million people are currently living with HIV, and some 20 million people have already died, with the worst of the epidemic centred on sub-Saharan Africa. But just as the spread of HIV has been greater than predicted, so too has been its impact on social capital, population structure and economic growth. Responding to AIDS on a scale commensurate with the epidemic is a global imperative, and the tools for an effective response are known. Nothing less than a sustained social mobilization is necessary to combat one of the most serious crises facing human development.**

**O**ver the 20 years since it was first identified, the HIV/AIDS epidemic has continued to exceed all expectations in the severity and scale of its impact. In 1991, the then Global Programme on AIDS at the World Health Organization forecast that by 2000 the cumulative global total of HIV infections in men, women and children would be 40 million<sup>1</sup>. In reality, this prediction has proved a serious underestimate.

An estimated 36 million people worldwide are currently living with HIV, and some 20 million people have already died, giving a cumulative total number of HIV infections of 56 million<sup>2</sup> — almost as many as the population of the United Kingdom. The worst of the epidemic, and the region where the reality has most out-stripped the predictions, is sub-Saharan Africa where, at the end of 2000, there were an estimated 25.3 million people living with HIV. Sub-Saharan Africa has accounted for around three-quarters of the global death toll.

Just as the spread of HIV has been greater than was predicted a decade ago, so too has been its demographic, social and economic impact. But the impact witnessed so far is only a fraction of the impact to come, given the rapid spread of HIV over the past 20 years together with the long lag time between infection and the onset of severe HIV-related disease and symptoms. The course of HIV disease in those already infected, even with the most optimistic scenarios about the efficacy and accessibility of treatment, will have a profound impact on future rates of life expectancy and economic growth. AIDS constitutes one of the most serious crises currently facing human development, and threatens to reverse progress in the most severely affected countries by decades.

And if the number of new infections continues at the current rate, even the most devastating impact that can be anticipated from current levels of infection will seem minor compared with that of the future. There were 5.3 million new HIV infections globally in 2000, and there is clear potential for further massive spread. Prevalence in the world's most populous nations remains relatively low, but in many of them the preconditions already exist for an epidemic that will be counted in the tens of millions of people affected, unless concerted and effective measures to turn the epidemic around are implemented without delay.

## An expanding epidemic

### High-income countries

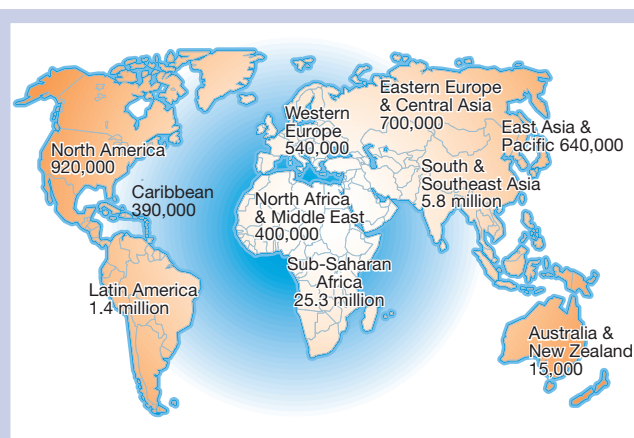
In the course of 2000, an estimated 30,000 adults and children became infected with HIV in Western Europe and 45,000 in North America, taking the total numbers living

with HIV in these regions combined to 1.46 million (Fig. 1). In these countries, the face of the epidemic has changed. Prevalence is increasing modestly. Persistent incidence at relatively low levels over the past decade has combined with longer survival. Highly active antiretroviral therapy (HAART) continues to have a favourable impact on disease progression and mortality in settings where it is available to people living with HIV, although the marked falls in mortality of 1996 and 1997 have levelled off in the past two years<sup>3-5</sup>.

In ethnic minority and poorer populations within high-income countries the beneficial effects of therapy are less pronounced. In addition, the benefits of effective prevention efforts are not felt to the same extent in minority and poorer populations. The proportion of AIDS incidence in these populations is increasing faster than in more affluent population sectors. Heterosexual sex now accounts for the largest single transmission category of AIDS cases in France<sup>6</sup>, and HIV incidence attributed to heterosexual sex with a person from a high-prevalence country has increased in a number of European countries<sup>5,7</sup>. In the United States, HIV/AIDS is also affecting minority populations disproportionately<sup>8</sup>, with disadvantaged young rural African-Americans one of the groups at high risk of HIV infection<sup>9</sup>.

HAART may also be one element in an increasing pattern of 'safe sex fatigue', leading to more risk-taking behaviour and HIV transmission especially among men who have sex with men. A study<sup>10</sup> among young men who have sex with men conducted in metropolitan areas in the United States found an HIV prevalence of 7.2%. Less than 20% of those infected knew of their infection before the study, and 41% reported unprotected anal sex during the past 6 months<sup>10</sup>. Surveys conducted among men who have sex with men in London indicate a rise in unsafe sex practices from 1996 to 1998<sup>11</sup> and an association between optimism about treatments and risk practices has been found in a number of gay communities<sup>12</sup>.

Prevention efforts have stalled in most industrialized countries, and the epidemic is moving slowly but apparently inexorably into the most vulnerable populations. But there are some successful examples that have defied this trend. Australia has a successful record in preventing an HIV epidemic in its injecting drug users, by acting, before there was marked incidence among injectors, to introduce needle and syringe exchange schemes and to support changed behavioural norms on needle sharing. The most recent national surveillance indicates HIV prevalence has been less than 0.5% in both men and women who identified themselves as injecting drug users at metropolitan sexual health centres from 1994 to 1999. HIV prevalence among



**Figure 1** Worldwide numbers of adults and children estimated to be living with HIV/AIDS at the end of 2000. (Source: ref. 2.)

people attending needle and syringe programmes has also remained low (less than 2%) except among men who identified themselves as homosexual<sup>13</sup>. This contrasts with HIV prevalence rates in injecting drug users of up to 18% in Chicago and 24–30% in New York<sup>14</sup>.

#### Eastern Europe

Eastern Europe has witnessed the greatest recent growth in the epidemic. At the start of January 1999, there were 10,000 reported HIV cases in the Russian Federation; now there are 70,000. More new infections occurred in 2000 than in all previous years combined (Fig. 2). Estimates of the actual numbers of people living with HIV have risen from 130,000 at the end of 1999 to 300,000 by the end of 2000. Four years ago the epidemic was reported from only a few cities, but by the end of 2000 there were cases reported in 82 out of the 89 regions of the Russian Federation<sup>15</sup>.

The epidemic in this region has been confined predominantly to injecting drug users, where HIV prevalence ranges from an estimated 0.2% to a high of 65% of needle and syringe exchange clients in Odessa, Ukraine. During 2000, new epidemics emerged in Uzbekistan and in Estonia, a country that reported ten times as many HIV cases in 2000 as in any previous year. These epidemics are likely to spread rapidly, given the high numbers of injecting drug users in countries experiencing economic and social upheaval (for example, more than 2 million estimated in the Russian Federation), the spread of HIV to the sexual partners of injecting drug users, and the high levels of sexually transmitted infections and sex work.

#### Sub-Saharan Africa

Sub-Saharan Africa is the region of the world where the epidemic has been worst and where the devastation of its impact is increasing. Average prevalence in sub-Saharan Africa is 8.8% in the adult population (15–49 years old); the region includes seven countries, all in the southern cone, with adult prevalence of 20% or more and a further nine countries over 10% (Fig. 3). The upper bound for prevalence is yet to be established, but Botswana has the highest rate so far with 36% adult prevalence, followed by Swaziland, Zimbabwe and Lesotho, which all have prevalence between 24 and 25%. Among the drivers of the epidemic in the southern cone has been the mobility that comes with a relatively extensive transport infrastructure, and work migration, especially of miners.

East Africa once had the highest infection rates on the continent but has now been overtaken by the southern cone. The prevalence rate among adults in Ethiopia and Kenya has reached double-digit figures. West Africa is only relatively less affected by HIV infection. Côte d'Ivoire ranks among the 15 worst-affected countries in the world. In Nigeria, adult prevalence of HIV is just over 5 per cent and has been rising slowly but steadily. Given that its population of

109 million is by far the region's largest, an increased epidemic in Nigeria could by itself add significantly to the continent's total.

Estimated new HIV infections in 2000 in sub-Saharan Africa numbered 3.8 million, a similar number to those infected in 1999. This total aggregates different situations across the region. In some districts and countries the epidemic continues to grow strongly, whereas in others the epidemic is long-standing and has already reached large numbers of people whose behaviour exposes them to HIV, leaving a smaller pool of people who can still be infected. In yet other countries, prevention efforts have successfully reduced the number of new infections.

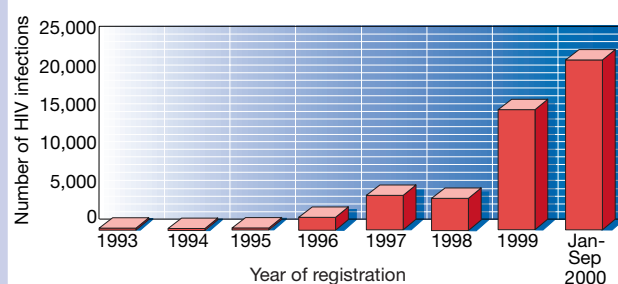
Effective prevention measures in some countries have enabled people to reduce their risk of exposure. The effects of behavioural change on HIV prevalence are clearest in relation to younger populations. In Zambia, for example, HIV prevalence for women under 20 attending antenatal clinics in Lusaka declined from 27% in 1993 to 17% by 1998; outside major urban areas it declined from 14% in 1994 to 6% in 1998<sup>16</sup>. In the Mbeya region of Tanzania, prevalence in 15–24-year-old attendees at antenatal clinics dropped from a peak of 20% in 1994 to 15% in 1999<sup>17</sup>. In Uganda, peak prevalence across the population of almost 14% in the early 1990s has since declined to around 8% and declining prevalence observed in pregnant women attending sentinel antenatal clinics is confirmed by rural cohort data<sup>18</sup>. In addition, some countries have good records of maintaining low prevalence. In Senegal, for example, infection rates have remained consistently below 2%, largely owing to the success of a nationwide campaign to modify sexual behaviour.

In general, the epidemic across sub-Saharan Africa has shifted from those at highest risk to a generalized epidemic, and from one concentrated in urban areas to one whose effects are felt almost as much in rural areas. Transmission continues to be dominated by heterosexual sex, together with a significant level of mother-to-child transmission. By the end of 2000 there were an estimated 1.1 million children under 15 living with HIV in sub-Saharan Africa, over 90% of whom acquired infection from their mother. As the epidemic has matured, the conditions that systematically generate vulnerability to HIV infection have also come into greater focus, including poverty, inequality and migration — the last of these being both labour-related and caused by conflict or civil strife<sup>19</sup>.

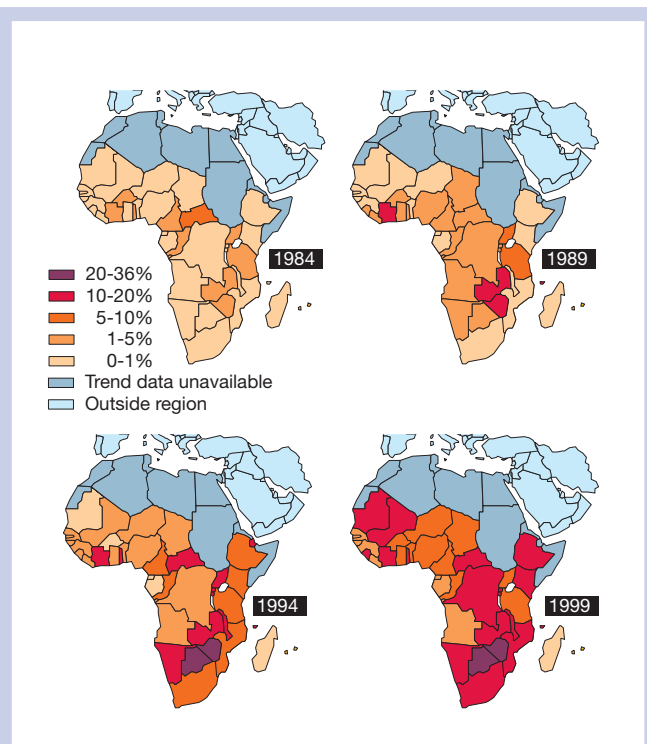
#### Asia, the Pacific and the Middle East

East Asia and the Pacific are still keeping HIV at bay, in regions that include some of the world's most populous countries. An estimated 640,000 people were living with HIV or AIDS at the end of 2000<sup>2</sup>. This represents just 0.07 per cent of the regions' adult population, as compared with the prevalence rate of 0.56 per cent in south and southeast Asia. However, there is undeniable potential for increasing spread in east Asia.

The HIV epidemic is linked closely to an epidemic of injecting drug use in many countries of the region<sup>20</sup>. In China, sentinel surveillance among injecting drug users detected no HIV infection in any of



**Figure 2** Annual number of newly registered HIV infections in Russia from 1993 until September 2000. (Source: Russian Federal AIDS Centre, Moscow.)



**Figure 3** Spread of HIV in Africa, 1984–1999. Estimated HIV prevalence rate in 15–49 year olds in sub-Saharan countries in 1984, 1989, 1994 and 1999. (Source: ref. 28.)

eight sentinel sites in 1995, but HIV was found in 17 of 19 sites in 1998. The highest prevalence among injecting drug users was 82% in Yining city in the western Xinjiang province<sup>21</sup>. Sentinel surveillance among sex workers in China shows little or sporadic condom use, and HIV prevalence ranging up to 6% in 1999<sup>22</sup>. Having practically eradicated sexually transmitted infections by the 1970s, China is now seeing a steep rise in the number of reported cases: from 5,800 in 1985 to over 836,000 in 1999<sup>23</sup>. This increase in sexually transmitted infections indicates potential for HIV spread in the future (Fig. 4).

In south and southeast Asia there were an estimated 700,000 new HIV infections in 2000, of which 64% were in men. This is consistent with risk behaviours in the region where men are the majority of injecting drug users and where they also drive the earliest wave of sexual transmission, much of it through commercial sex and some through sex between men.

The epidemic in India is very diverse, which is not unexpected in a subcontinent whose population of a billion is substantially greater than the whole of Africa's. In southern states there are epidemics driven largely by heterosexual sex, with adult prevalence of 2% in Tamil Nadu<sup>22</sup>. Among pregnant women, prevalence is above 1% in Maharashtra, Andhra Pradesh, Karnataka and Tamil Nadu. Meanwhile, in the northeastern states of Manipur and Nagaland there are significant epidemics among injecting drug users, albeit with much smaller absolute numbers<sup>24</sup>.

Recent national surveillance in Bangladesh has shown continuing low levels of HIV infection in a number of high-risk groups, ranging from 2.5% among injecting drug users, to 1.5% among brothel-based sex workers, to 0.2% in men who have sex with men<sup>25</sup>. However, the same report found high levels of risk behaviour in these groups indicating substantial potential for future spread of HIV.

The most-affected countries in southeast Asia, with adult prevalence over 2%, are Thailand, Cambodia and Myanmar. Thailand and Myanmar have both seen epidemics simultaneously among injecting drug users, men who have sex with men and heterosexuals with a high turnover of partners, whereas Cambodia's epidemic has been driven

mainly by male norms of premarital and extramarital sex, mostly with women who are paid. The success of HIV prevention in Thailand, especially interventions directed at commercial sex workers and clients, has been well documented, and sustained<sup>26</sup>, even in the more severely affected northern provinces<sup>27</sup>. However, there has been recent public disquiet as to their future sustainability with reduced public health investment in HIV prevention following the Asian economic downturn.

In the Middle East and North Africa, recent evidence indicates that new infections are on the rise in this region but from a base that until now has remained very low. For example, localized studies in one area of Algeria show rates of around 1% in pregnant women attending an antenatal clinic<sup>2</sup>, and surveillance sites in Sudan, especially in the south, indicate that HIV is spreading among the general population.

#### Latin America and the Caribbean

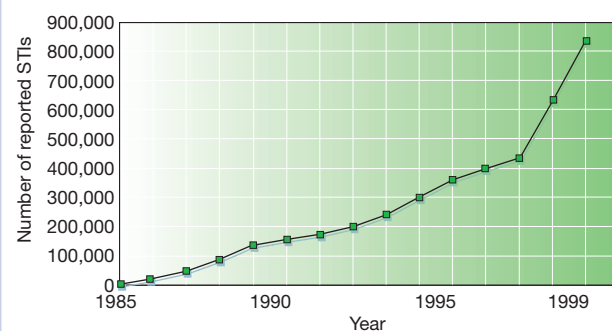
The epidemic in Latin America is a complex patchwork of HIV spread through sex between men, sex between men and women, and injecting drug use. When the region is taken as a whole, incidence is not increasing, with an estimated 150,000 adults and children becoming infected during 2000. However, this is not a uniform pattern: several Central American countries have seen a recent rise in levels of infection. In contrast, Brazil has demonstrated lower than anticipated HIV incidence (presentation by the Ministry of Health, Brazil, at Okinawa International Conference on Infectious Diseases, 7–8 December 2000), with a combination of concerted prevention efforts, including those involving men who have sex with men, and a rights-based approach to antiretroviral provision with access to therapy enshrined in the constitution.

AIDS mortality has declined, especially in Brazil with the impact of antiretroviral therapy, and consequently Latin and Central America has experienced a rise in prevalence. By the end of 2000, some 1.4 million adults and children were estimated to be living with HIV or AIDS, compared with 1.3 million at the end of 1999.

Rates are generally high in the Caribbean where heterosexual spread of HIV is the predominant mode of transmission. Five countries have HIV prevalence rates of over 1% of the adult population. In Haiti by the end of 1999, the prevalence rate among the adult population exceeded 5%, the only country with such rates outside the African continent<sup>28</sup>.

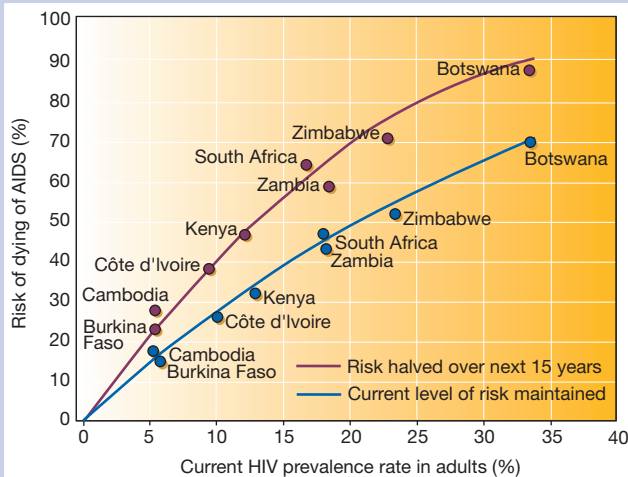
#### Society-wide impacts of HIV

The epidemiological impact of HIV only begins to convey the full impact of HIV/AIDS on the current and future capacity of societies to sustain the dignity and security of human life. AIDS magnifies its impact into the future because it erodes social capital. Because AIDS affects primarily young adults and in many cases initially spreads more extensively in the more mobile, wealthier and better educated parts of populations, its effects ramify across all social and economic sectors. It is only when HIV becomes entrenched that it reproduces



**Figure 4** Reported sexually transmitted infections in China, 1985 to 1999. (Source: ref. 22.)





**Figure 5** Lifetime risk of dying of AIDS for 15-year-old boys from selected countries, assuming unchanged or halved risk of becoming infected with HIV. (Source: ref. 28.)

patterns of wider social vulnerability. Hence the positive association between higher educational attainment and likelihood of HIV infection, which is characteristic early in the epidemic and which has been reported in Africa and Latin America, reverses as the epidemic matures to a persistent association between HIV and lower educational levels<sup>29</sup>.

At the macro-level, HIV has sustained impacts on development, whether measured in terms of life expectancy or in terms of economic growth. Here, we will mainly discuss its impact in sub-Saharan Africa, because that is where it has been most sustained and most studied. However, the conclusions drawn about the nature of the impact hold equally true elsewhere in the world for current or future HIV epidemics at scales comparable with that in sub-Saharan Africa.

In the worst affected countries, HIV is wiping out decades of gains in life expectancy. In southern Africa, average life expectancy rose from 44 years in the early 1950s to 59 years by the late 1980s, but over the coming decades the impact of HIV will return life expectancy to under 45 years<sup>30</sup>.

More than 20 million people have died worldwide since the beginning of the epidemic, three-quarters of them in sub-Saharan Africa. Globally, HIV/AIDS is now well established in the list of the top five leading causes of death. It is surpassed only by disease groups such as ischaemic heart disease, cerebrovascular disease and lower respiratory infections—whose predominance is largely because they are typical causes of death for old people<sup>31</sup>. In sub-Saharan Africa, AIDS is responsible for one in five deaths, twice as many as for the second leading cause of death.

The demographic impact of AIDS is unique for two reasons. First, unlike most other causes of death, AIDS deaths will continue to rise in the coming years as a result of infections that have already occurred. Second, HIV infection is highest in young women and men in their most productive years, including those in the best-educated and skilled sectors of populations, as well as women of child-bearing age, together with attendant transmission to children. In the worst affected countries, in 20 years time the standard population pyramid will have turned upside down, with more adults in their 60s and 70s than those in their 40s and 50s<sup>32</sup>.

Current prevalence data do not convey the full picture facing individuals in populations with high HIV prevalence. Because prevalence is a measure of current infection levels among living individuals, it does not capture infections among those who have already died or who have not yet become infected but will be in the future. On the basis of current incidence and mortality patterns, it is possible to estimate the lifetime risks of contracting HIV and dying from AIDS faced by young people embarking on the sexually active phase of their

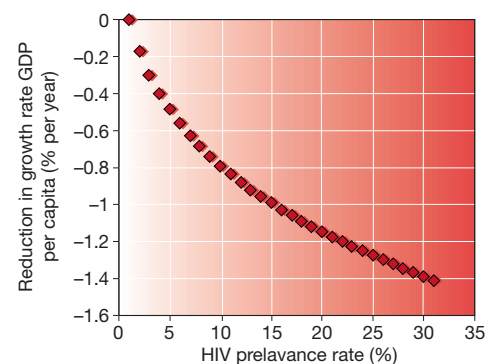
lives. In a country such as South Africa or Zambia, where prevalence in 2000 has reached about 20%, a 15-year old teenager faces a lifetime risk of HIV infection and of death from AIDS of over 50%, unless the current rate of new infections drops dramatically (Fig. 5).

The impact on rates of economic growth in developing countries is equally marked and there is a direct relationship between the extent of HIV prevalence and the severity of reduction in growth of gross domestic product (GDP; Fig. 6). Measures of per capita GDP in fact underestimate the human impact of AIDS, as AIDS kills people as well as economic activity. The cumulative impact of HIV on the total size of economies is thus even greater. By the beginning of the next decade, South Africa, which represents 40% of sub-Saharan Africa's economic output, is facing a GDP that will be 17% lower than it would have been without AIDS<sup>33</sup>. In settings where subsistence agriculture predominates, measured economic productivity only scratches the surface of the total impact of HIV on livelihoods. For example, AIDS hits the long-term capacity for agricultural production, as livestock is often sold to pay funeral expenses, or orphaned children lack the skills to look after livestock in their care<sup>34,35</sup>.

The immediate impact of AIDS is felt most acutely in households where one or more members is infected with HIV (Fig. 7). In South Africa, households will have on average 13 per cent less to spend per person by 2010 than they would if there were no HIV epidemic. In Côte d'Ivoire, the household impact of HIV/AIDS not only reverses the capacity to accumulate savings, but also reduces household consumption. In addition to affecting income, with lower earning capacity and productivity, AIDS generates greater medical, funeral and legal costs, and has long-term impact on the capacity of households to stay together. This is manifest in the cumulative number of children orphaned by AIDS, which now totals 13.2 million.

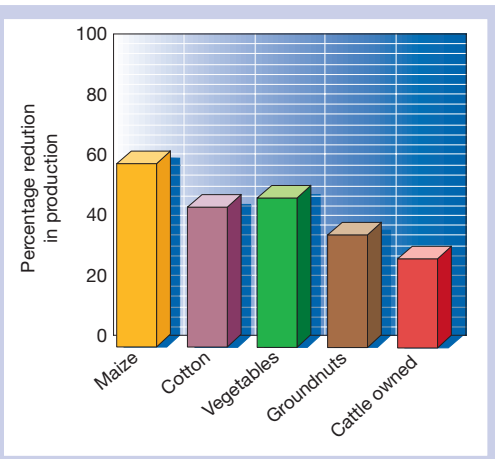
Household impact is one of the points at which AIDS and poverty demonstrate their inter-twined relationship. AIDS exacerbates and prolongs poverty in every context. For example, although Botswana has the highest per capita incomes in sub-Saharan Africa, a proportion of its population lives in poverty. In 1996 this proportion had fallen to 38% from 49% a decade earlier, but in another decade, AIDS will result in a return to 45% (ref. 36). In poorer households, AIDS takes a greater proportion of available expenditure, and limits access to everything from healthcare to food.

In education, AIDS has a negative impact both on the supply of teachers and on the capacity of children to continue in school. In Zambia, for example, children of AIDS-affected families are likely to drop out of school in urban areas because their carers do not have the funds available for school fees, and in rural areas they may also be required to work in the fields<sup>37</sup>. Across sub-Saharan Africa an estimated 860,000 children lost their teachers to AIDS in 1999, and in some countries many more teachers die than retire, and the



**Figure 6** Growth impact of HIV. Graph shows the relationship between HIV prevalence and reduction in growth rate measured as per capita GDP. Data obtained from 80 developing countries for the period 1990–1997. (Source: ref. 50.)

**Figure 7** Reduction in production in a Zimbabwean household with an AIDS death. (Source: ref. 51).



teaching force is being depleted almost as rapidly as new teachers can be trained<sup>38</sup>.

Across the labour force generally, AIDS is significantly affecting the workforce, including its most skilled elements. By 2020, the workforce in 15 countries analysed by the International Labour Organization will be 24 million people fewer because of AIDS<sup>39</sup>. The workforce costs of AIDS are substantial and include lost productivity due to health-related absence and funeral attendance, higher payments for insurance and medical care, and hiring and retraining costs. In some cases as many as one-third of employees may be infected, as found in a study of miners in their late 20s and 30s in southern Africa<sup>40</sup>. In these contexts, spending by employers to reduce the HIV-related risk of their employees is highly cost-effective. For example, a comprehensive intervention addressing South African miners and their community context—including the women selling them sex—found that its impact in averting infections saved the company 25 times the cost of the intervention<sup>41</sup>.

The breadth of the impact of AIDS has transformed it from an issue that can be understood in relation to discrete sectors of human activity to one that needs to be appreciated as a fundamental issue of human security. These are the terms in which it was twice debated at the United Nations Security Council in the course of 2000. The security implications of AIDS range from the higher levels of infection found among military personnel<sup>42</sup>, the destabilizing effects on urban and national security of cohorts of young people without hope and often without parents as a result of HIV<sup>43</sup>, to the geopolitical effects of depleted national infrastructures and human resources in the worst-affected countries.

### Effective responses make a difference

Against this background it is important to remember that an ever-growing AIDS epidemic is not inevitable. Even within severely affected regions, some nations or communities within nations have managed to sustain low incidence and others with large established epidemics have managed to achieve reduced incidence in key populations.

The lessons of the past 20 years of response to the epidemic demonstrate the preconditions of effective control: unified national planning, the application of proven strategies on a scale commensurate with the epidemic, access to essential prevention and care commodities, and a public environment conducive to the social inclusion of people with HIV and those most at risk. Although it is universally the case that HIV transmission results from very particular risk behaviours, the capacity of individuals to control their level of risk, that is, their HIV vulnerability, varies widely according to the social environment in which the risk occurs. In turn, HIV vulnerability and the impact of HIV are proportionally related, as an increasing impact depletes social stability, support and cohesion.

These lessons apply whether an epidemic is in its infancy, where securing control is easiest, or in established and generalized

epidemics, where the task of reversing the epidemic is hardest. But despite what is known about how to be effective in preventing the further spread and ameliorating the impact of HIV/AIDS, there is a huge gap between the current level of activity and what is needed. The urgent task now is to ensure that the lessons of effective practice are applied at a society-wide scale in every setting, guided both by local context and by scientific evidence.

Because HIV transmission occurs mainly in intimate contexts, HIV infection raises issues of individual and collective identity. The temptation therefore exists to respond to HIV as a platform for constructing ideal identity, at the expense of responses guided by the real world of actual behaviour and risk practice, and pragmatic measures to reduce vulnerability and alleviate the impact of HIV. The need for interventions to be guided by evidence rather than ideology is therefore paramount. This is nowhere more evident than in relation to sex education for children—the precondition for their adopting life-long protective strategies. Evidence indicates that far from encouraging sexual activity, childhood sex education results in higher levels of risk reduction, such as delaying the onset of sexual activity, and higher levels of condom use<sup>44</sup>. Equally, a pragmatic response to harm minimization for injecting drug users needs to attend to the increasing evidence that needle and syringe exchange is effective in preventing HIV transmission, and has no other harmful effects<sup>45</sup>.

The most appropriate targets for intervention cannot be prescribed centrally, although they can follow globally appropriate principles. Effective responses to HIV need to constitute a variety of community-level interventions responsive to particular contexts of risk and vulnerability. In Managua, Nicaragua, motels are the principal venue for sexual activity outside the home, including a large proportion of commercial and extramarital sex, so actually handing condoms to motel room users results in greater use than does distribution of HIV education materials<sup>46</sup>. The prevention and treatment of sexually transmitted infections is a major adjunct to HIV prevention<sup>47</sup>. Voluntary counselling and testing has been shown to be both effective in prevention<sup>48</sup> and cost-effective<sup>49</sup>, in addition to its pivotal role as the gateway to care for people living with HIV and in promoting societal and personal openness about AIDS.

There are three over-riding universal principles that underpin the myriad successful local responses to HIV. First is the inclusion of people living with HIV centrally in the response, as a source both of creativity in devising solutions and of accountability in focusing on the realities of the epidemic. Second, prevention and care need to be pursued in synergy, as experience has shown that the most effective responses to the epidemic integrate education, prevention and care strategies as interdependent elements of the response. Third, stigma associated with HIV continues to constitute a major barrier to effective action. A concerted effort against stigma will not only improve the quality of life of people living with HIV and those who are most vulnerable to infection, but also meet one of the necessary conditions of a full-scale response to the epidemic. When these principles are applied to local responses, and when the political leadership exists to proliferate local responses on a national scale, the epidemic can be reversed.

### Conclusion: social mobilization as the fundamental response

The great historical advances in public health have all involved a radical reconceptualization of the social as well as the medical technologies associated with particular diseases. Recognizing and responding to the issues of social organization that are specific to particular disease entities have been the key to effective responses to major epidemics (for example, quarantine between towns in the case of plague, and urban sanitation in the case of cholera). Smallpox vaccination seems the ultimate 'magic bullet'—an example of a purely medical technology resulting in the total eradication of disease. But even the response to smallpox cannot escape its social determinants, such as the observation of natural immunity in cow-herding communities, the power relations that enabled vaccine

testing, and the 200 years of social, political and economic organization required for vaccination to occur globally.

HIV/AIDS has its particular biological and social characteristics that dictate the shape of an effective response: its impact is greatest among young adults; the virus is transmitted through intimate behaviours; its impact ramifies across every field of human endeavour; infection may remain invisible for many years; and overcoming the stigmatization of people with HIV infection, or thought to be at heightened risk, is a precondition for explicit action against the disease.

Biomedical intervention in the form of antiretroviral therapy has substantially reduced the morbidity and mortality associated with HIV for those populations able to afford access, and has had a significant impact on preventing mother-to-child transmission. However, when looked at globally, these successes have been relevant to only a small proportion of the world's population affected by HIV. Other potential biomedical interventions in the form of effective topical microbicides, or vaccines, have to date proved elusive.

The types of effective responses to HIV epidemics around the world derive from these specific biological and social characteristics. What is required is nothing less than a sustainable social mobilization. Its key elements are the involvement of affected communities, including individuals who are infected; restructuring of global finance flows so that the essential commodities required for the response can be made available universally; and systematically targeting social exclusion.

Responding to HIV/AIDS on a scale commensurate with the epidemic is a global imperative. The task has barely begun, but at least we are at the end of the beginning, with the needs recognized together with the proven elements of an effective response. The imperative now shifts to garnering the requisite global, national and community leadership that will be the only basis on which the total social mobilization against AIDS can be sustained. □

- Global Programme on AIDS. *Current and Future Dimensions of the HIV/AIDS Pandemic* (World Health Organization, Geneva, 1991).
- Joint United Nations Programme on HIV/AIDS and World Health Organization. *AIDS Epidemic Update: December 2000* (UNAIDS, Geneva, 2000).
- The CASCADE Collaboration. Survival after introduction of HAART in people with known duration of HIV-1 infection. *Lancet* **355**, 1158–1159 (2000).
- Centers for Disease Control and Prevention. HIV/AIDS Surveillance Rep. 12, 35 (2000).
- Institut de Veille Sanitaire Surveillance du Sida en France. Bulletin Épidémiologique Hebdomadaire **38** (2000).
- Cazein, F. *et al.* Tendances récentes de l'épidémie de Sida. *Sida, VIH et MST, Etat des Données en 2000* (Institut de Veille Sanitaire, Saint-Maurice, 2000).
- Public Health Laboratory Service. Diagnosed HIV infected people in England and Wales—projections for the next four years. *Communic. Dis. Rep.* **10**, 453–454 (2000).
- Smith, D. K. *et al.* HIV/AIDS among African Americans: progress or progression? *AIDS* **14**, 1237–1248 (2000).
- Centers for Disease Control and Prevention. Cluster of HIV-infected adolescents and young adults—Mississippi, 1999. *Morbidity Mort. Weekly Rep.* **49**, 861–864 (2000).
- Valleroy, L. A. *et al.* HIV prevalence and associated risks in young men who have sex with men. *J. Am. Med. Assoc.* **284**, 198–204 (2000).
- Dodds, J. P. *et al.* Increasing high risk sexual behaviour among homosexual men in London 1996–8. *Br. Med. J.* **320**, 1510–1511 (2000).
- Bolding, G. *et al.* International differences among gay men in HIV optimism and sexual risk behaviour—a report from London, Melbourne, Sydney and Vancouver. Abstract LbPp105, XIIIth International AIDS Conference, Durban, South Africa, 9–14 July 2000. <<http://www.iac2000.org>>
- National Centre in HIV Epidemiology and Clinical Research. *HIV/AIDS 2000 Hepatitis C and Sexually Transmissible Infections in Australia Annual Surveillance Report 2000* (National Centre in HIV Epidemiology and Clinical Research, The University of New South Wales, Sydney, 2000).
- Monterroso, E. R. *et al.* Prevention of HIV infection in street-recruited injection drug users. *J. Acquir. Immune Defic. Syndr.* **25**, 63–70 (2000).
- Dehne, K. L., Pokrovskiy, V., Kobyscha, Y. & Schwartz, B. Update on the epidemics of HIV and other sexually transmitted infections in the newly independent states of the former Soviet Union. *AIDS* **14**(Suppl.), S75–S84 (2000).
- Joint United Nations Programme on HIV/AIDS. *AIDS in Africa, Country by Country* (UNAIDS, Geneva, 2000).
- Jordan-Harder, B., Koshima, Y. A., Pervillac, C. & Vogel, U. *Hope for Tanzania: Lessons Learned from a Decade of Comprehensive AIDS Control in Mbeya Region* (GTZ Eschborn, Germany, and Ministry of Health, Dar-es-Salaam, 2000).
- Kamali, A. *et al.* Seven-year trends in HIV-1 infection rates, and changes in sexual behaviour, among adults in rural Uganda. *AIDS* **14**, 427–434 (2000).
- Carael, M., Buve, A. & Awusabo-Asare, K. The making of HIV epidemics: what are the driving forces? *AIDS* **11**(Suppl.), S23–S31 (1997).
- Beyrer, C. *et al.* Overland heroin trafficking routes and HIV-1 spread in south and south-east Asia. *AIDS* **14**, 75–83 (2000).
- Qu, S., Sun, X. & Zheng, X. National sentinel surveillance of HIV infection in China from 1995 to 1998. Abstract MoPeC2379, XIIIth International AIDS Conference, Durban, South Africa, 9–14 July 2000. <<http://www.iac2000.org>>
- Joint United Nations Programme on HIV/AIDS. *Monitoring the AIDS Pandemic. The Status and Trends of the HIV/AIDS Epidemics in the World. Provisional Report 5–7 July 2000* (UNAIDS Geneva, 2000).
- Chen, X.-S., Gong, X.-D., Liang, G.-J. & Zhang, G.-C. Epidemiologic trends of sexually transmitted diseases in China. *Sex. Transm. Dis.* **27**, 138–142 (2000).
- Shaukat, M. HIV sentinel surveillance in India. Abstract MoPeC2416, XIIIth International AIDS Conference, Durban, South Africa, 9–14 July 2000. <<http://www.iac2000.org>>
- AIDS and STD Control Programme. *Report on the Sero-Surveillance and Behavioural Surveillance on STD and AIDS in Bangladesh 1998–1999* (Ministry of Health & Family Welfare, Government of the People's Republic of Bangladesh, 2000).
- Khamboonruang, C. *et al.* Sustained low HIV-1 incidence among cohorts of young northern Thai adults: confirmation of Thailand's success in HIV prevention. Abstract ThOr671, XIIIth International AIDS Conference, Durban, South Africa, 9–14 July 2000. <<http://www.iac2000.org>>
- Kilmarx, P. H. *et al.* Explosive spread and effective control of human immunodeficiency virus in northernmost Thailand: the epidemic in Chiang Rai province, 1988–99. *AIDS* **14**, 2731–2740 (2000).
- Joint United Nations Programme on HIV/AIDS. *Report on the Global HIV/AIDS Epidemic, June 2000* (UNAIDS, Geneva, 2000).
- Vandemoortele, J. & Delamónica, E. The “education vaccine” against HIV. *Curr. Issues Comp. Educ.* **3** <<http://www.tc.columbia.edu/cice/vol03nr1/jvedart1.htm>> (2000).
- Population Division, United Nations, and Joint United Nations Programme on HIV/AIDS. *The Demographic Impact of HIV/AIDS. Report of Technical Meeting, New York, 10 November 1998* (United Nations, New York, 1999).
- World Health Organization. *World Health Report 2000* (WHO, Geneva, 2000).
- Schwartzländer, B., Garnett, G., Walker, N. & Anderson, R. AIDS in a new millennium. *Science* **289**, 64–67 (2000).
- Arndt, C. & Lewis, J. The macro implications of HIV/AIDS in South Africa: a preliminary assessment. *S. Afr. J. Econ.* **68**(3), 1–32 (2000).
- Engh, I. D., Stloukal, L. & du Guerny, J. *HIV/AIDS in Namibia: The Impact on the Livestock Sector* (Food and Agriculture Organization, Rome, 2000).
- Ayieko, M. From single parents to child headed households: the case of children orphaned by AIDS in Kismu and Siaya districts. Study No. 7 (UNDP, New York, 1998).
- Botswana Institute for Development Policy Analysis. *Macroeconomic Impacts of the HIV/AIDS Epidemic in Botswana* (BIDPA, Botswana, 2000).
- Nampunya-Serpell, N. Social and economic risk factors for HIV/AIDS—affected families in Zambia. AIDS and Economics Symposium, Durban, 7–8 July 2000. <<http://www.iaen.org/conferences/durbansym/papers/67nampunya.pdf>>
- United Nations Children's Fund. *The Progress of Nations 2000* (UNICEF, New York, 2000).
- International Labour Organization. *HIV/AIDS A Threat to Decent Work, Productivity and Development* (ILO, Geneva, 2000).
- Evian, C. Anonymous HIV surveillance: a powerful tool for advocacy and for managing HIV in the workplace. Abstract MoPeC2365, XIIIth International AIDS Conference, Durban, South Africa, 9–14 July 2000. <<http://www.iac2000.org>>
- Michael, K. A critique of best practices: company activities on HIV/AIDS in South Africa. Abstract MoPeD2852, XIIIth International AIDS Conference, Durban, South Africa, 9–14 July 2000. <<http://www.iac2000.org>>
- Sena, A. *et al.* Trends of gonorrhea and chlamydial infection during 1985–1996 among active-duty soldiers at a United States Army installation. *Clin. Infect. Dis.* **30**, 742–748 (2000).
- Mendoka Mj, M. Speech to the 18th Commission for Human Settlement Conference, Gigiri, Kenya. Reported in the *East African Standard*, 17 February 2001.
- Joint United Nations Programme on HIV/AIDS. *Force for Change: World AIDS Campaign with Young People* (UNAIDS, Geneva, 1998).
- Blumenthal, R. N., Kral, A. H., Erringer, E. A. & Edlin, B. R. The effect of syringe exchange use on high-risk injection drug users: a cohort study. *AIDS* **14**, 605–611 (2000).
- Egger, M. *et al.* Promotion of condom use in a high-risk setting in Nicaragua: a randomised controlled trial. *Lancet* **355**, 2101–2105 (2000).
- Grosskurth, H., Gray, R., Hayes, R., Mabey, D. & Wawer, M. Control of sexually transmitted diseases for HIV-1 prevention: understanding the implications of the Mwanza and Rakai trials. *Lancet* **355**, 1981–1987 (2000).
- The Voluntary HIV-1 Counseling and Testing Efficacy Study Group. Efficacy of voluntary HIV-1 counselling and testing in individuals and couples in Kenya, Tanzania, and Trinidad: a randomised trial. *Lancet* **356**, 103–112 (2000).
- Sweat M *et al.* Cost-effectiveness of voluntary HIV-1 counselling and testing in reducing sexual transmission of HIV-1 in Kenya and Tanzania. *Lancet* **356**, 113–121 (2000).
- Bonnel, R. *Economic Analysis of HIV/AIDS. Background Paper for the African Development Forum, September 2000* (The World Bank, Washington DC, 2000).
- Bollinger, L. & Stover, J. *The Economic Impact of AIDS* (The Futures Group International, Washington DC, 1999).



# The dynamics of CD4<sup>+</sup> T-cell depletion in HIV disease

Joseph M. McCune

Gladstone Institute of Virology and Immunology, Departments of Medicine and Microbiology and Immunology, University of California at San Francisco, PO Box 419100, San Francisco, California 94141-9100, USA (e-mail: mmccune@gladstone.ucsf.edu)

**The size and composition of the CD4<sup>+</sup> T-cell population is regulated by balanced proliferation of progenitor cells and death of mature progeny. After infection with the human immunodeficiency virus, this homeostasis is often disturbed and CD4<sup>+</sup> T cells are instead depleted. Such depletion cannot result simply from accelerated destruction of mature CD4<sup>+</sup> T cells — sources of T-cell production must also fail. Ironically, this failure may be precipitated by physiological mechanisms designed to maintain homeostasis in the face of accelerated T-cell loss.**

In most parts of the world, most of the time, infection with the human immunodeficiency virus (HIV) leads to death. Shortly after this syndrome was first defined, it became clear that the virus could replicate within human CD4<sup>+</sup> T cells *in vitro*, that the viral envelope protein could bind to CD4, and that circulating CD4<sup>+</sup> T cells decreased in number as disease progressed<sup>1,2</sup>. It was also apparent that CD4<sup>+</sup> T 'helper' cells were crucial in coordinating cellular and humoral immune responses against exogenous antigens. As a result of these observations, it was not difficult to imagine that HIV-associated immunodeficiency was due to virally mediated destruction of CD4<sup>+</sup> T cells.

Now, many years later, much remains left to the imagination. We still do not know how, *in vivo*, the virus destroys CD4<sup>+</sup> T cells or whether, in quantitative terms, cell loss is due to direct destruction by virus or to other indirect means<sup>3</sup>. This ignorance, arising in large part because it is difficult to study the immune system in living human beings, hinders the discovery and development of effective vaccines and therapies. Several hypotheses have been proposed to explain the loss of CD4<sup>+</sup> T cells, some of which seem to be diametrically opposed. Here, I summarize these hypotheses and incorporate them within a single model for disease progression.

## The evidence for CD4<sup>+</sup> T-cell depletion

When a patient presents with anaemia, medical students reflexively run a drill: the red blood-cell count may be low because the cells have been destroyed, lost from the intravascular space, or not produced in sufficient quantities to meet demand. Each of these scenarios carries a distinct prognosis and a defined set of treatments, and people with anaemia usually do not improve unless the correct diagnosis is made.

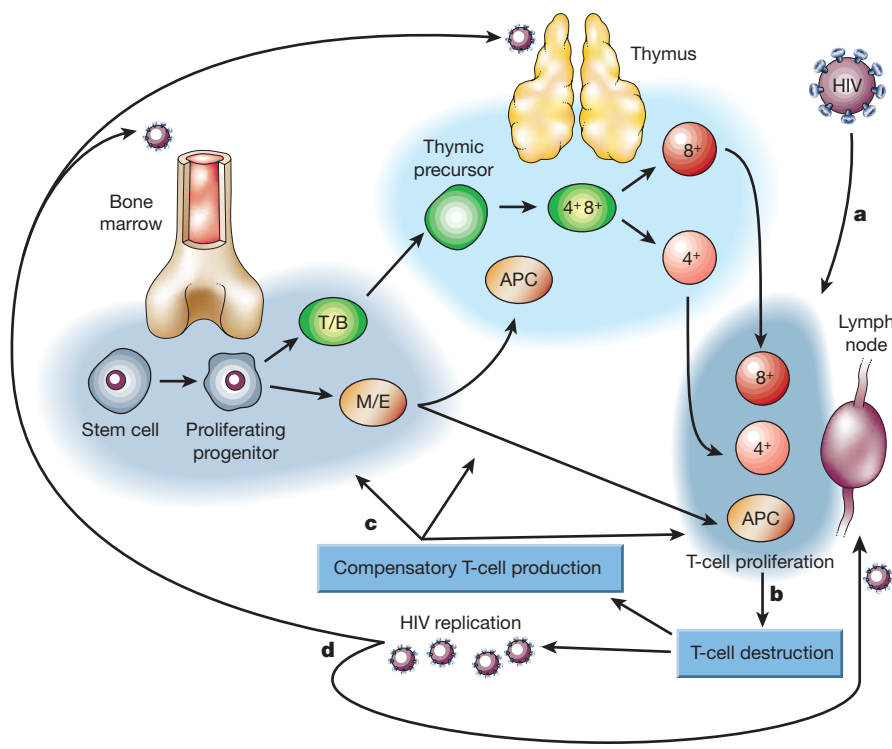
The analysis of T lymphopenia is less precise. There is a growing realization that the T-lymphocyte compartment is comprised of multiple subpopulations<sup>4</sup> and that there are inadequate means by which to monitor the relative growth, death and movement of these subpopulations in an individual patient over time<sup>5</sup>. Definition of CD4<sup>+</sup> T-cell loss in HIV disease thus remains unclear.

Most researchers would agree that HIV infection results in progressive loss of CD4<sup>+</sup> T cells from the circulation as well as depletion of CD4<sup>+</sup> T cells from total body stores.

Quantitative estimates, using a variety of assumptions, indicate that the normal young (<30 year old) adult harbours about  $2 \times 10^{11}$  mature CD4<sup>+</sup> T cells<sup>6</sup>. In the HIV-infected patient, this total number is halved by the time the peripheral blood CD4<sup>+</sup> T-cell count falls to 200 cells  $\mu\text{l}^{-1}$  (refs 6, 7). In more advanced disease, destruction of parenchymal lymphoid spaces is so extensive that enumeration of the total body CD4<sup>+</sup> T-cell count has not even been attempted. With disease progression, there is also a decrease in the proportion of quiescent naive (CD45RA<sup>+</sup>CD62L<sup>+</sup>) T cells and an increase in the proportion of activated memory/effector (CD45RO<sup>+</sup>) T cells; concomitantly, the T-cell receptor (TCR) repertoire is both perturbed and restricted<sup>8</sup>. Among those cells that persist, many may be dysfunctional<sup>9</sup>. In sum, HIV induces both quantitative and qualitative defects in the CD4<sup>+</sup> T-cell compartment and, as imprecise as it may be, the circulating CD4<sup>+</sup> T-cell count continues to be one of the best surrogate markers by which to gauge prognosis late in infection and to trigger treatment interventions.

## Possible causes of T-cell depletion

Several mechanisms have been proposed to explain HIV-mediated depletion of CD4<sup>+</sup> T cells. All are based on sound experimental observations and some are intuitively obvious. Not surprisingly, these mechanisms recapitulate the differential diagnosis of anaemia. Thus, total body CD4<sup>+</sup> T cells may be depleted in absolute number because they are destroyed or because their production is impaired. In addition, the fraction of circulating cells may decrease (giving the appearance of loss) if viral infection results in their redistribution out of the intravascular space and into the confines of lymphoid organs. As in the case of anaemia, the balance of destruction and production is one that can be tipped by multiple mechanisms (Table 1). It is possible, for instance, that CD4<sup>+</sup> T-cell depletion is related directly to the virally mediated destruction of infected cells. Alternatively, physiological responses to HIV infection might initiate events that result in the destruction of uninfected cells. In either case, loss of mature cells should be compensated by increased production of new cells and mature CD4<sup>+</sup> T-cell depletion should occur only if cells lost in the periphery cannot be replaced. The devastating feature of HIV infection is that the virus can have direct and indirect pathogenic effects on both mature CD4<sup>+</sup> T cells and on the progenitor cells from which they arise. In each of the following sections,



**Figure 1** Accelerated T-cell destruction leads to impaired production. The figure shows a model of HIV-induced disease progression and CD4<sup>+</sup> T-cell depletion. Depicted are three interlinked organs of the haematolymphoid system. The bone marrow contains multilineage haematopoietic stem cells and high proliferative potential cells necessary for generation of the lymphoid (T/B) and myeloerythroid (M/E) lineages. The thymus supports T-cell differentiation from intrathymic T progenitor cells (thymic precursors) and selects mature CD3<sup>+</sup>CD4<sup>+</sup> (4<sup>+</sup>) and CD3<sup>+</sup>CD8<sup>+</sup> (8<sup>+</sup>) thymocytes from CD3<sup>int</sup>CD4<sup>+</sup>CD8<sup>+</sup> (4<sup>+</sup>8<sup>+</sup>) cortical thymocytes by mechanisms that rely in part on bone marrow-derived myeloid antigen-presenting cells (APCs). Peripheral lymphoid organs are where antigen-dependent (CD3<sup>+</sup>CD4<sup>+</sup> and CD3<sup>+</sup>CD8<sup>+</sup>) T- and B-cell responses are coordinated by bone marrow-derived APCs. CD4 and chemokine co-receptors for HIV may be expressed on cells at multiple layers of this hierarchy. The structure and function of these organs may be compromised by a variety of direct and/or indirect effects of HIV infection<sup>43,44</sup>. In the young adult, bone marrow-derived T progenitor cells continue to mature, albeit at a reduced rate, through the thymus and the peripheral lymphoid system is continuously seeded with naive T cells bearing a maximally diverse TCR repertoire. When HIV is introduced into the body (stage a), it is probably concentrated within draining lymph nodes and presented as an antigen, resulting in enhanced movement of T cells into nodes and vigorous T-cell proliferation. Immune activation caused by HIV will be associated, by direct and indirect means, with accelerated destruction of T cells (stage b). To compensate for peripheral lymphoid depletion, 'sensor cells'<sup>68</sup> in lymphoid organs may produce positive regulatory signals (for example, IL-7) which promote enhanced T-cell production at multiple levels of the haematopoietic tree (stage c). These signals may in turn accelerate virus replication and the emergence of highly cytopathic variants, leading to the destruction of key progenitor cells in the bone marrow, the thymus and/or the peripheral lymphoid system (stage d). Impaired production of new cells from these organs would then result in collapse of the immune system and conversion of a high turnover state to one of low turnover.

these various mechanisms are discussed in isolation. More accurately, they should be viewed as interacting mechanisms that operate in concert.

#### Accelerated destruction of mature CD4<sup>+</sup> T cells

Early experiments done with laboratory-adapted HIV isolates in tissue culture revealed a cytopathic virus with exquisite tropism for CD4<sup>+</sup> T cells. It was reasonable to predict that CD4<sup>+</sup> T cells should be equally susceptible to HIV-induced death *in vivo*. Multiple experiments done in the late 1980s and early 1990s provided support for this view and the hypothesis of 'accelerated destruction' received indirect experimental validation in a series of influential experiments by Ho and Shaw and their colleagues in 1995<sup>10,11</sup>. The provision of potent (protease inhibitor-containing) antiretroviral medications to patients with advanced HIV disease caused the viral load to drop and the CD4<sup>+</sup> T-cell count to rise. Given reasonable and largely accepted assumptions about T-cell distribution and assuming that antiretroviral therapy does not alter the production rate of T cells, these data were interpreted to mean that, before therapy, continuous rounds of *de novo* infection sustained the viral load and that as many as  $2 \times 10^9$  infected CD4<sup>+</sup> T cells were destroyed per day. By extension, HIV disease was a 'high turnover' state marked by 'virologic mayhem'<sup>12</sup>,

accelerated destruction of mature CD4<sup>+</sup> T cells, and eventual 'exhaustion' of the immune system<sup>10,11</sup>.

The corresponding image of a sink with an ever-expanding drain was a powerful stimulus for further experimentation. In HIV-infected human subjects, quantitative image analysis revealed decreased numbers of CD4<sup>+</sup> T cells and increased levels of cellular proliferation and apoptosis in lymphoid tissue<sup>6</sup>. In rhesus macaques infected with simian immunodeficiency virus (SIV), turnover rates were found to increase substantially in multiple populations of cells, including memory and naive T cells, natural killer cells and B cells<sup>13,14</sup>. Analysis of T-cell turnover in humans with HIV infection confirmed estimates from these previous studies: the fraction of dividing CD4<sup>+</sup> T cells in untreated HIV disease can be elevated 2–3-fold<sup>15,16</sup>, with most proliferation concentrated in the CD45RO<sup>+</sup> population of memory/effector CD4<sup>+</sup> T cells<sup>15,17</sup>. Given the fact that cell division must be balanced by cell death at steady state, these data are consistent with the interpretation that HIV infection accelerates both the production and the destruction of CD4<sup>+</sup> T cells.

There are, however, some problems with this scenario. First, there seem to be many more cells dividing and dying than there are cells infected<sup>18,19</sup>. Second, even though the fraction of dividing cells may be high, the absolute number of such cells may actually decrease when

the CD4<sup>+</sup> T-cell pool size shrinks (Box 1). Third, CD8<sup>+</sup> T-cell division and death rates are at least as high as those of CD4<sup>+</sup> T cells<sup>14,16</sup>. Finally, destruction alone cannot explain the pronounced changes that occur in the composition and function of the CD4<sup>+</sup> T-cell compartment. These problems prompted further thought about interactions that might occur between HIV and CD4<sup>+</sup> cells (T and non-T cells) within the confines of haematolymphoid organs.

#### Altered movement leading to the appearance of loss

The immune system comprises a series of highly organized parenchymal spaces, to and through which T cells and antigens must move and meet. Linked in series by a network of blood vessels and lymphatics, these lymphoid organs provide residence for the vast majority of all CD4<sup>+</sup> and CD8<sup>+</sup> T cells. Movement of cells from one location to another is directed in part by homing receptors that mediate tissue-specific interactions with endothelial cells<sup>20</sup>. Naive (CD45RA<sup>+</sup>CD62L<sup>+</sup>) T cells recirculate in an L-selectin (CD62L)-dependent manner from blood to lymph node and then back to blood. In addition, memory/effector (CD45RO<sup>+</sup>) T cells move through non-lymphoid tissues such as the liver and lung, thereby increasing the chances of encountering foreign microorganisms<sup>21</sup>. When antigen-presenting cells (APCs) initiate immune responses within these lymphoid organs, CD4<sup>+</sup> T cells are stimulated and retained within them while activated CD8<sup>+</sup> T cells move into the circulation. The net outcome of such differential movement is a transient decrease in the total lymphocyte count, a CD8 lymphocytosis, and inversion of the CD4/CD8 ratio<sup>22,23</sup>.

Given this physiological framework, it is predictable that the normal immune response against HIV might prompt differential movement of T cells out of the circulation and into fixed lymphoid organs<sup>24–27</sup>. This shift would not change the absolute number of total body CD4<sup>+</sup> T cells but could have pronounced effects on the number and proportion of T-cell subpopulations in the circulation. Changes of this type have indeed been observed in the context of acute and chronic infections with HIV or SIV<sup>6,28,29</sup> and may be even further augmented by upregulation of CD62L on CD4<sup>+</sup> T cells after infection with HIV<sup>30</sup>. Reciprocal changes may then occur after the initiation of effective antiretroviral therapy: a drop in viral load may dampen immune activation and permit sequestered cells to redistribute back into the peripheral circulation<sup>31–34</sup>. Consistent with this possibility are observations that reductions in viral load are associated with decreased levels of adhesion molecules (for example, VCAM-1 and ICAM-1), which normally mediate lymphocyte sequestration into lymphoid tissue, and with decreased size and cell content of superficial nodes<sup>35</sup>. Whether or not cell destruction occurs, these shifts in cell movement could account for many of the changes observed in the composition of the CD4<sup>+</sup> T-cell compartment in untreated and treated patients.

#### Chronic activation and T-cell death

T cells do die, however, in the setting of HIV disease and much of this death may occur in uninfected cells as a by-product of HIV infection elsewhere. According to this view<sup>24,36–38</sup>, HIV disease is typified by a state of chronic activation, driven in part by the antigenic stimulus of HIV and in part by antigen-independent mechanisms — for example, cytokines released by APCs and activated T cells. Multiple bursts of activated cells, spread throughout the body, would be characterized by APC-mediated activation of resting lymphocytes, cytokine-driven expansion of responding cells, and contraction of the responding population by activation-induced cell death (AICD)<sup>39</sup>. If APCs are infected by or otherwise carrying HIV, antigen-specific cell activation could support virus dissemination to responding CD4<sup>+</sup> T cells, irrespective of their TCR specificity<sup>40</sup>. Reciprocally, the virus may be spread upon activation of non-productively infected CD4<sup>+</sup> memory T cells in the context of immune responses to HIV or other antigens. During the asymptomatic phase of infection, when the fraction of infected cells is much lower than the

**Table 1 Causes of T-cell depletion: destruction of mature CD4<sup>+</sup> T cells**

- Direct destruction of infected cells
  - Envelope-mediated apoptosis
  - Vpr-induced G2 arrest and apoptosis
  - Disruption of cell membrane integrity/syncytia formation
  - Accumulation of unintegrated viral DNA
- Indirect induction of death in uninfected cells
  - Cytotoxicity by HIV-specific cytolytic T cells or by natural killer cells
  - Autoimmune reactions of a humoral or cellular nature
  - Incorporation into syncytia by neighbouring infected cells
  - Triggering of apoptosis upon cell activation or cross-linking of CD4-bound gp120.
  - Enhanced HIV transmission and/or apoptosis following interaction with nearby infected antigen-presenting cell

Each of these mechanisms is supported by considerable *in vitro* and some *in vivo* data<sup>1,275–80</sup> and all may be important. It has not been possible, however, to quantitate the relative contribution of each to progressive CD4<sup>+</sup> T-cell depletion *in vivo*.

fraction of activated cells<sup>19</sup>, these bursts would predictably continue in a local, recurrent and asynchronous fashion, and CD4<sup>+</sup> T-cell depletion might be driven by several mechanisms. First, relentless activation of naive T cells into the activated/memory pool may not be fully compensated by replenishment of new naive cells from the thymus (see below) or by the generation of viable memory cells<sup>36</sup>. Alternatively, chronic stimulation of 'resting' T cells might have a negative effect on the homeostatic regeneration of these cells. The outcome would be manifest by a higher fraction of dividing CD4<sup>+</sup> T cells as well as by progressive contraction of the CD4<sup>+</sup> TCR repertoire and anergy (unresponsiveness to antigen).

The relevance of this process to CD4<sup>+</sup> T-cell depletion is underscored by the observation that disease progression is associated with immune activation, and vice versa<sup>41</sup>. The repercussions of this interplay are particularly evident in patient populations affected by other chronic infectious diseases, for example, those caused by helminthic parasites, tuberculosis and/or malaria. Even before HIV infection supervenes, chronic immune activation induced by these organisms is associated with many of the immunological signs of HIV disease (for example, decreased CD4<sup>+</sup> and increased CD8<sup>+</sup> T-cell counts, elevated levels of activated CD4<sup>+</sup> and CD8<sup>+</sup> T cells, marked increases in T-cell apoptosis, and impairment of T-cell functional responses to antigen)<sup>42</sup>. Given concurrent infection with HIV, a bad situation is likely to deteriorate quickly.

#### Impaired production of new T cells

The above mechanisms focus on mature CD4<sup>+</sup> T cells, but these cells are often derived from early progenitors that may also express CD4. Such progenitors, including multilineage and lineage-restricted haematopoietic progenitor cells, are uniquely endowed with the capacity to persist with long half-lives and to generate large numbers of differentiated progeny rapidly upon stimulation. If these cells are destroyed or rendered nonfunctional, mature progeny could not be made.

Evidence for suppression of multilineage and lineage-specific haematopoiesis has been available since the beginning of the AIDS epidemic<sup>43,44</sup>. When late-stage patients initially presented with opportunistic infections, they were not just lymphopenic, but anaemic, neutropenic and thrombocytopenic as well. These findings led to multiple diagnostic bone-marrow biopsies, the results of which were frequently abnormal. Microscopic examination revealed hypercellularity or hypocellularity, plasmacytosis, myeloid or erythroid dysplasia, and a variety of other pathological changes<sup>43</sup>. Phenotypic and functional analysis of bone-marrow progenitor cells showed a decrease in the number of lineage-restricted colony-forming units and, in some but not all instances, infection and/or apoptotic death of CD34<sup>+</sup> progenitors. Although the mechanisms associated with such cytopenias remain obscure (Table 2), they are often reversed upon the provision of effective antiretroviral therapy<sup>15,16,45,46</sup>.

The thymus, housing many CD4<sup>+</sup> T cells in varying stages of maturation, is another critical target organ for HIV infection<sup>47</sup>. Examination of paediatric and adult specimens has revealed



thymocyte depletion, loss of corticomedullary demarcation and development of thymic medullary B-cell follicles. These changes are associated with immunohistochemical visualization of HIV structural proteins within thymocytes and are evidence of viral replication. Studies in the SCID-hu Thy/Liv mouse have shown that such destruction is most rapidly effected by CXCR4-utilizing (X4) isolates of HIV which are tropic and cytolytic for CXCR4<sup>+</sup>CD3<sup>+</sup>CD4<sup>+</sup>CD8<sup>-</sup> intrathymic T progenitor cells<sup>48,49</sup>. By contrast, CCR5-utilizing (R5) isolates infect intrathymic myeloid cells and thymocyte subpopulations that are more mature, and result in thymocyte depletion only after much longer periods of time<sup>49</sup>. Although it has proven difficult to study the thymus in HIV-infected humans, the frequency of circulating CD4<sup>+</sup> and CD8<sup>+</sup> naive (CD45RA<sup>+</sup>CD62L<sup>+</sup>) T cells has been found to decrease as disease progresses<sup>50,51</sup>. In addition, cells bearing TCR excision circles (TRECs; markers of recent intrathymic TCR rearrangement) also decrease in frequency with age and as a function of HIV disease progression<sup>52–54</sup>. Reciprocally, signs of thymopoiesis return after treatment of some individuals with effective antiretroviral therapy<sup>17,34,52,55,56</sup>, particularly if they are younger and have evidence of abundant thymus by computed tomography<sup>57</sup>.

Finally, peripheral lymphoid organs undergo marked alterations after HIV infection. These changes include accumulation of virus on, and eventual destruction of, the follicular dendritic cell network, decompartmentalization, and depletion of both the CD4<sup>+</sup> and CD8<sup>+</sup> T-cell populations<sup>6,25</sup>. Antigen-dependent and antigen-independent expansion is less likely to occur either because the stromal microenvironment necessary for antigen presentation is disrupted or because the memory/effector T cells responsive to antigen are depleted, or both.

Thus, HIV infection leads to profound disruption of the bone marrow, thymus and peripheral lymphoid organs and, where measurable, quantitative and qualitative defects in important T progenitor cells. With these cells eliminated or no longer functional, the immune system cannot be sustained.

### Resolving the conundrum

Each of the above mechanisms of CD4<sup>+</sup> T-cell depletion has been championed by thoughtful investigators with logical arguments and strong supporting data. In some instances, one mechanism may in fact be dominant over others. Consider, for example, the following two extremes of infection *in utero* and infection late in life.

Before the sixteenth week of gestation, few T cells reside in the periphery of the human fetus; most CD4<sup>+</sup> cells are found instead within the fetal liver, bone marrow or thymus, and many of these are multilineage or lineage-restricted haematopoietic progenitor cells<sup>47</sup>. If HIV crosses the placenta and infects these organs, the immune system may not form properly. Such regenerative failure probably underlies paediatric 'thymic dysfunction'<sup>58</sup> and may be responsible for the rapid disease progression observed in 20–30% of HIV-infected children<sup>59</sup>.

Rapid disease progression has also been observed when HIV infection occurs late in life<sup>60</sup>. Because older individuals normally have minimal residual thymic tissue and diminished peripheral lymphoid reserve<sup>61</sup>, their ability to regenerate T cells from central sources is compromised severely. Immunodeficiency is instead most easily attributed to accelerated destruction of mature CD4<sup>+</sup> T cells, either

directly or through the ever-increasing stress of chronic immune activation and its effects on peripheral homeostasis.

Worldwide, these two extremes represent only a tiny fraction of the HIV-infected population. Far more HIV-positive individuals are young adults with an immune system that is fully developed and endowed with considerable regenerative reserve, including the presence of a functioning thymus<sup>52–54,62,63</sup>. In most of these individuals, the disease course is marked by chronic progression over a period of 5–10 years. This relatively long duration is explained most satisfactorily by a time-dependent, sequential juxtaposition of multiple mechanisms of CD4<sup>+</sup> T-cell depletion (Fig. 1). In this model of disease progression, all of the above mechanisms of CD4<sup>+</sup> T-cell depletion may have interconnecting roles.

As the initiating step in this sequence (stage a in Fig. 1), HIV must be inoculated into the body and carried to draining lymph nodes. There, it will be presented as a collection of antigenic epitopes, triggering vigorous proliferation of antigen-specific T and B cells. This response will prompt differential movement of antigen-specific CD4<sup>+</sup> T cells into lymph nodes, exponential proliferation of T effector cells and high levels of AICD (stage b, Fig. 1). Within a period of several weeks, the entering bolus of infecting virions may be neutralized by antibody and/or by cytotoxic T lymphocytes, selecting resistant variants for outgrowth over time<sup>64</sup>. Such variants might then initiate chronic and repeating cycles of *de novo* antigen presentation, immune cell activation, neutralization and escape. Overall, presentation of HIV as an antigen will result in chronic immune activation accompanied by AICD and, among those cells that survive, an increased degree of anergy.

Unlike other antigens, HIV also kills. It is clear that HIV can infect APCs such as macrophages<sup>65</sup> and become sequestered within emigrating dendritic cells<sup>66</sup>. If and when these cells present antigen to CD4<sup>+</sup> T cells, the virus can be transmitted across close cell–cell contacts in a highly efficient manner<sup>40</sup>. Because these APCs present epitopes that are representative of environmental antigens (some of which are likely to be associated with invading microorganisms), this process will result in selective removal of TCR specificities that would be best maintained and, over time, accelerated destruction of mature CD4<sup>+</sup> T cells (stage b, Fig. 1).

Predictably, the haematopoietic system will not stand still in the face of such destruction. This system exists under clear (if poorly understood) homeostatic control and, in non-T-cell lineages, destruction of mature elements is compensated by *de novo* production from progenitor cells. For example, loss of red blood cells or platelets stimulates the peripheral production of erythropoietin or thrombopoietin, which circulate back to the bone marrow to inhibit programmed cell death and to stimulate the proliferation of early erythroid or megakaryocytic progenitors. Similarly, accelerated destruction of mature T cells might stimulate *de novo* and accelerated production of T cells (stage c, Fig. 1). If so, and if the rate of production keeps pace with the rate of destruction, the CD4<sup>+</sup> T-cell count should be sustained. For physiological immune function to remain intact, it is also important that compensatory feedback on the T-cell lineage generates a diverse and normal complement of CD4<sup>+</sup> T cells. For this to occur, it may be important to summon residual thymic reserves. Do adaptations of this sort occur in HIV-infected people?

The phenomenon of 'thymic rebound' has been noted previously in uninfected adults, often in association with peripheral T lymphopenia induced by myeloablation<sup>62</sup>. Similarly, HIV-infected adults with declining CD4<sup>+</sup> T-cell counts (in the range of 300–500 cells  $\mu\text{L}^{-1}$ ) also have more thymic tissue (as viewed by computed tomography) than age-matched, uninfected controls<sup>63</sup>. This tissue seems to be functioning, as evidenced by the presence of higher circulating levels of naive (CD45RA<sup>+</sup>CD62L<sup>+</sup>) CD4<sup>+</sup> T cells containing a higher frequency of TRECs (J. Harris and J. M. McCune, unpublished observations, 2000) and a predictably long half-life<sup>17</sup>. Patients with abundant thymus also show a faster return of naive CD4<sup>+</sup> T cells after antiretroviral therapy<sup>57</sup>. Such thymic rebound has been ascribed to

**Table 2 Causes of T-cell depletion: impaired T-cell production**

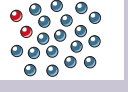
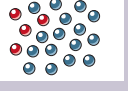
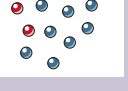
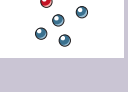
- Direct effects of virus
  - Infection-mediated death of progenitor cells
  - Destruction of the supporting stromal network required for multilineage or lineage-restricted haematopoiesis
- Indirect effects of virus
  - Cytokine dysfunction
  - Opportunistic infections of bone marrow (for example, by cytomegalovirus or *Mycobacterium avium intracellulare*)
  - HIV-induced apoptosis
  - Infiltrating malignancies
  - Myelotoxic effects of drugs
  - Deficiencies of vitamins and of other essential factors

## Box 1

Considerations of T-cell turnover *in vivo*

In the context of the CD4<sup>+</sup> T-cell compartment, the term 'turnover' refers to the flux of newly divided cells moving through the system in a unit period of time. If the T-cell pool size remains in steady state, the addition of a newly produced cell must be counterbalanced by the loss of another cell. The fractional replacement rate  $k$  refers to the fraction of cells newly produced in a unit period of time. This can be estimated on a static basis by measuring the number of cells in S phase (for example, by staining for Ki67) and on a kinetic basis by measuring the number of dividing cells (for example, those that incorporate pyrimidine nucleosides such as bromodeoxyuridine or tritiated thymidine, or that incorporate stable isotope-enriched nucleoside precursors such as deuterated glucose)<sup>81</sup>. The absolute turnover rate is then calculated as the product of  $k$  and the pool size. In previous studies of T-cell turnover in HIV or SIV disease, a high  $k$ -value (measured using bromodeoxyuridine or deuterated glucose) has at times been interpreted to reflect a high degree of T-cell production and destruction<sup>13,14</sup>. But as the examples in the table below illustrate, this conclusion may not always be correct. If the pool size is normally 20 cells, a  $k$ -value of 0.1 per day results in an absolute turnover rate of 2 cells per day (top row). A higher  $k$ -value of 0.2 per day, however, can reflect an absolute turnover rate that is high, normal or low depending on whether the pool size itself is reduced (lower rows). In late stages of HIV disease, when the total body CD4<sup>+</sup> T-cell pool is small, a high  $k$ -value may therefore be associated with a low absolute turnover rate.

Box 1 Table Pool size and replacement rate affect T-cell turnover

|                                                                                     | Pool size     | $k$           | Absolute turnover rate |
|-------------------------------------------------------------------------------------|---------------|---------------|------------------------|
|  | Normal<br>20  | Normal<br>0.1 | Normal<br>2            |
|  | Normal<br>20  | High<br>0.2   | High<br>4              |
|  | Reduced<br>10 | High<br>0.2   | Normal<br>2            |
|  | Reduced<br>5  | High<br>0.2   | Low<br>1               |

either increased peripheral secretion of a positive regulator (for example, interleukin (IL)-2, IL-7, *flt3L*, stem-cell factor, growth hormone, insulin-like growth factor-1 or thyroid hormone) or decreased secretion of a negative regulator (for example, testosterone, oestrogen or corticosteroids), or both<sup>62</sup>. Among these, studies indicate that IL-7 may be a principal contributor to T-cell homeostasis in the general setting of lymphopenia<sup>67</sup> and in the specific instance of HIV disease<sup>68,69</sup>. Thus, in cross-sectional and longitudinal studies of HIV-positive individuals at varying stages of disease, falling CD4<sup>+</sup> T-cell counts are associated with increased circulating levels of IL-7; reciprocally, IL-7 levels fall when CD4<sup>+</sup> T-cell counts increase after antiretroviral treatment. Because IL-7 is known to inhibit programmed cell death and to stimulate the proliferation of both intrathymic T progenitor cells and more mature T-cell subpopulations<sup>67</sup>, such increases in IL-7 may result in increased T-cell production.

Compensatory feedback control over T-cell production might suffice in the short term, but it does not solve and may eventually

exacerbate HIV-induced lymphopenia. Ironically, high levels of IL-7-driven T-cell division create a fertile substrate for viral spread: not only will the virus replicate more efficiently in dividing cells, but IL-7 also acts as a cofactor in the transactivation of the viral long terminal repeat<sup>70</sup>. Increased levels of viral replication may then have several effects (stage d in Fig. 1). First, more progenitor cells may be infected and destroyed. Second, the generation of highly cytopathic (for example, X4) variants may accelerate destruction of important CXCR4<sup>+</sup>CD4<sup>+</sup> progenitor cells in the bone marrow, thymus and peripheral lymphoid organs.

A final twist to this sequence of events arises from the observation that, in patients who do not respond to protease-inhibitor therapy, high levels of protease inhibitor-resistant virus can coexist with persistently high levels of circulating CD4<sup>+</sup> T cells<sup>71,72</sup>. These T cells show reduced signs of activation, turn over slowly, and are enriched for those with a naive phenotype<sup>73</sup>. When tested *in vitro*, the protease inhibitor-resistant viruses from these patients replicate as well as wild-type virus in mature, mitogen-activated CD4<sup>+</sup> T cells. They do not, however, replicate efficiently in thymocytes<sup>74</sup>. Such selective tropism, sparing the thymus and possibly also CD4<sup>+</sup> T memory cells, may represent one instance in which accelerated destruction of mature cells can be counterbalanced by compensatory production of new cells.

### Clinical considerations

To summarize, CD4<sup>+</sup> T cells may be lost in HIV disease because the virus delivers two blows: it destroys mature effector T cells and it reduces the ability of the body to replace these cells from immature progenitors. Accelerated destruction may occur because HIV is an effective antigen, prompting redistribution of responding T cells into lymphoid organs, initiation of a vigorous and chronic proliferative response, and induction of cell death by direct and indirect means. Destruction of mature T cells should then trigger compensatory feedback signals that promote T lymphopoiesis, until such time as the organs of T-cell production are themselves destroyed. Regenerative failure would at that point lead to collapse of the immune system.

This model incorporates a series of lesions that explain the dynamics of HIV-associated CD4<sup>+</sup> T-cell depletion. It also enables a more physiological analysis of HIV-induced pathology. Rapid disease progression in adults may be associated with several key variables such as exogenous (non-HIV-related) antigens that drive the activation of the immune system and the diminished haematological reserves of old age. By contrast, the beneficial effects of treatment intervention may be most apparent in those with residual haematological reserves (that is, young people with a functioning thymus and/or with little prior destruction of the associated peripheral CD4<sup>+</sup> T-cell compartment). Such individuals would predictably show better responses both to antiretroviral therapies (provided either continuously or in an interrupted manner) and to therapeutic immunization. If so, then additional therapies may be required for those who lack sufficient reserves and/or who continue to harbour a high burden of exogenous antigen. Further analysis of these possibilities may provide insight into effective therapies for pathophysiologically distinct subgroups of individuals with HIV disease. □

- Levy, J. A. Pathogenesis of human immunodeficiency virus infection. *Microbiol. Rev.* **57**, 183–289 (1993).
- D'Souza, M. P. & Fauci, A. S. in *Textbook of AIDS Medicine* (eds Merigan, T. C., Bartlett, J. G. & Bolognesi, D.) 59–85 (Williams and Wilkins, Baltimore, 1999).
- Zinkernagel, R. M. Are HIV-specific CTL responses salutary or pathogenic? *Curr. Opin. Immunol.* **7**, 462–470 (1995).
- Mitra, D. K. *et al.* Differential representations of memory T cell subsets are characteristic of polarized immunity in leprosy and atopic diseases. *Int. Immunol.* **11**, 1801–1810 (1999).
- Westermann, J. & Pabst, R. Lymphocyte subsets in the blood: a diagnostic window of the lymphoid system? *Immunol. Today* **9**, 43–45 (1990).
- Haase, A. T. Population biology of HIV-1 infection: viral and CD4<sup>+</sup> T cell demographics and dynamics in lymphatic tissues. *Annu. Rev. Immunol.* **17**, 625–656 (1999).
- Rosok, B. I. *et al.* Reduced CD4 cell counts in blood do not reflect CD4 cell depletion in tonsillar tissue in asymptomatic HIV-1 infection. *AIDS* **10**, F35–F38 (1996).
- Gorochov, G. *et al.* Perturbation of CD4<sup>+</sup> and CD8<sup>+</sup> T-cell repertoires during progression to AIDS and regulation of the CD4<sup>+</sup> repertoire during antiviral therapy. *Nature Med.* **4**, 215–221 (1998).
- Clerici, M. & Shearer, G. A Th1 to Th2 switch is a critical step in the aetiology of HIV infection. *Immunol. Today* **14**, 107–111 (1993).

10. Ho, D. D. *et al.* Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature* **373**, 123–126 (1995).
11. Wei, X. *et al.* Viral dynamics in human immunodeficiency virus type 1 infection. *Nature* **373**, 117–122 (1995).
12. Wain-Hobson, S. Virological mayhem. *Nature* **373**, 102 (1995).
13. Nowak, M. A. *et al.* Viral dynamics of primary viremia and antiretroviral therapy in simian immunodeficiency virus infection. *J. Virol.* **71**, 7518–7525 (1997).
14. Mohri, H., Bonhoeffer, S., Monard, S., Perelson, A. S. & Ho, D. D. Rapid turnover of T lymphocytes in SIV-infected rhesus macaques. *Science* **279**, 1223–1227 (1998).
15. Fleury, S. *et al.* Long-term kinetics of T cell production in HIV-infected subjects treated with highly active antiretroviral therapy. *Proc. Natl Acad. Sci. USA* **97**, 5393–5398 (2000).
16. Hellerstein, M. *et al.* Directly measured kinetics of circulating T lymphocytes in normal and HIV-1-infected humans. *Nature Med.* **5**, 83–89 (1999).
17. McCune, J. M. *et al.* Factors influencing T cell turnover in HIV-1-seropositive patients. *J. Clin. Invest.* **105**, R1–R9 (2000).
18. Harper, M. E., Marselle, L. M., Gallo, R. C. & Wong-Stall, F. Detection of lymphocytes expressing human T-lymphotropic virus type III in lymph nodes and peripheral blood from infected individuals by *in situ* hybridization. *Proc. Natl Acad. Sci. USA* **83**, 772–776 (1986).
19. Chun, T. W. *et al.* Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. *Nature* **387**, 183–198 (1997).
20. Butcher, E. C. & Picker, L. J. Lymphocyte homing and homeostasis. *Science* **276**, 60–66 (1996).
21. Mackay, C. R., Marston, W. L. & Dudler, L. Naive and memory T cells show distinct pathways of lymphocyte recirculation. *J. Exp. Med.* **171**, 801–817 (1990).
22. Bishop, D. K., Ferguson, R. M. & Orosz, C. G. Differential distribution of antigen-specific helper T cells and cytotoxic T cell after antigenic stimulation in vivo. *J. Immunol.* **144**, 1153–1160 (1990).
23. Bujdosó, R., Young, P., Hopkins, J., Allen, D. & McConnell, I. Non-random migration of CD4 and CD8 T cells: changes in the CD4:CD8 ratio and interleukin 2 responsiveness of effluent lymph node cells following in vivo antigen challenge. *Eur. J. Immunol.* **19**, 1779–1784 (1989).
24. Grossman, Z. & Herberman, R. B. T-cell homeostasis in HIV infection is neither failing nor blind: modified cell counts reflect an adaptive response of the host. *Nature Med.* **3**, 486–490 (1997).
25. Rosenberg, Y. J. & Janossy, G. The importance of lymphocyte trafficking in regulating blood lymphocyte levels during HIV and SIV infections. *Semin. Immunol.* **11**, 139–154 (1999).
26. Kirschner, D., Webb, G. F. & Cloyd, M. Model of HIV-1 disease progression based on virus-induced lymph node homing and homing-induced apoptosis of CD4<sup>+</sup> lymphocytes. *J. Acquir. Immune Defic. Syndr.* **24**, 352–362 (2000).
27. Hengel, R. L., Jones, B. M., Kennedy, M. S., Hubbard, M. R. & McDougal, J. S. Markers of lymphocyte homing distinguish CD4 T cell subsets that turn over in response to HIV-1 infection in humans. *J. Immunol.* **163**, 3539–3548 (1999).
28. Rosenberg, Y. J. *et al.* Decline in the CD4<sup>+</sup> lymphocyte population in the blood of SIV-infected macaques is not reflected in lymph nodes. *AIDS Res. Hum. Retroviruses* **9**, 639–646 (1993).
29. Grossman, Z., Bentwich, Z. & Herberman, R. B. From HIV infection to AIDS: are the manifestations of effective immune resistance misinterpreted? *Clin. Immunol. Immunopathol.* **69**, 123–125 (1993).
30. Wang, L., Chen, J. J., Gelman, B. B., König, R. & Cloyd, M. W. A novel mechanism of CD4 lymphocyte depletion involves effects of HIV on resting lymphocytes: induction of lymph node homing and apoptosis upon secondary signaling through homing receptors. *J. Immunol.* **162**, 268–276 (1999).
31. Mosier, D. E. HIV results in the frame: CD4<sup>+</sup> cell turnover. *Nature* **375**, 193–194 (1995).
32. Sprent, J. & Tough, D. HIV results in the frame: CD4<sup>+</sup> cell turnover. *Nature* **375**, 194 (1995).
33. Dimitrov, D. S. & Martin, M. A. HIV results in the frame: CD4<sup>+</sup> cell turnover. *Nature* **375**, 194–195 (1995).
34. Pakker, N. G. *et al.* Biphasic kinetics of peripheral blood T cells after triple combination therapy in HIV-1 infection: a composite of redistribution and proliferation. *Nature Med.* **4**, 208–214 (1998).
35. Bucy, R. P. *et al.* Initial increase in blood CD4<sup>+</sup> lymphocytes after HIV antiretroviral therapy reflects redistribution from lymphoid tissues. *J. Clin. Invest.* **103**, 1391–1398 (1999).
36. Hazenberg, M. D., Hamann, D., Schuitemaker, H. & Miedema, F. T cell depletion in HIV-1 infection: how CD4<sup>+</sup> T cells go out of stock. *Nature Immunol.* **1**, 285–289 (2000).
37. Grossman, Z., Feinberg, M. B. & Paul, W. E. Multiple modes of cellular activation and virus transmission in HIV infection: a role for chronically and latently infected cells in sustaining viral replication. *Proc. Natl Acad. Sci. USA* **95**, 6314–6319 (1998).
38. Anderson, R. W., Ascher, M. S. & Sheppard, H. W. Direct HIV cytopathicity cannot account for CD4 decline in AIDS in the presence of homeostasis: a worst-case dynamic analysis. *J. Acquir. Immune Defic. Syndr.* **17**, 245–252 (1998).
39. Krammer, P. H. CD95's deadly mission in the immune system. *Nature* **407**, 789–795 (2000).
40. Pope, M. *et al.* Conjugates of dendritic cells and memory T lymphocytes from skin facilitate productive infection with HIV-1. *Cell* **78**, 389–398 (1994).
41. Bentwich, Z., Kalinkovich, A., Weisman, Z. & Grossman, Z. Immune activation in the context of HIV infection. *Clin. Exp. Immunol.* **111**, 1–2 (1998).
42. Borkow, G. *et al.* Chronic immune activation associated with intestinal helminth infections results in impaired signal transduction and anergy. *J. Clin. Invest.* **106**, 1053–1060 (2000).
43. McCune, J. M. & Kaneshima, H. in *Human Hematopoiesis in SCID Mice* (eds Roncarolo, M.-G., Peault, B. & Namikawa, R.) 129–156 (Landes Publishing Company, Austin, TX, 1995).
44. Moses, A., Nelson, J. & Bagby, G. The influence of human immunodeficiency virus-1 on hematopoiesis. *Blood* **91**, 1479–1495 (1998).
45. Huang, S. S. *et al.* Reversal of HIV-1-associated hematosuppression by effective antiretroviral therapy. *Clin. Infect. Dis.* **30**, 504–510 (2000).
46. Fleury, S. *et al.* Limited CD4<sup>+</sup> T-cell renewal in early HIV-1 infection: effect of highly active antiretroviral therapy. *Nature Med.* **4**, 794–801 (1998).
47. McCune, J. M. HIV-1: the infective process in vivo. *Cell* **64**, 351–363 (1991).
48. Su, L. *et al.* HIV-1 induced thymocyte depletion is associated with indirect cytopathicity and infection of progenitor cells in vivo. *Immunity* **2**, 25–36 (1995).
49. Berkowitz, R. D. *et al.* R5 and X4 strains of HIV-1 exhibit differential tropism and pathogenesis in vivo. *J. Virol.* **72**, 10108–10117 (1998).
50. Rabin, R. L. *et al.* Altered representation of naive and memory CD8 T cell subsets in HIV-infected children. *J. Clin. Invest.* **95**, 2054–2060 (1995).
51. Roederer, M. *et al.* CD8 naive T cell counts decrease progressively in HIV-infected adults. *J. Clin. Invest.* **95**, 2061–2066 (1995).
52. Douek, D. C. *et al.* Changes in thymic function with age and during the treatment of HIV infection. *Nature* **396**, 690–695 (1999).
53. Poulin, J.-F. *et al.* Direct evidence for thymic function in adult humans. *J. Exp. Med.* **190**, 479–486 (1999).
54. Zhang, L. *et al.* Measuring recent thymic emigrants in blood of normal and HIV-1-infected individuals before and after effective therapy. *J. Exp. Med.* **190**, 725–732 (1999).
55. Autran, B. *et al.* Positive effects of combined antiretroviral therapy on CD4<sup>+</sup> T cell homeostasis and function in advanced HIV disease. *Science* **277**, 112–116 (1997).
56. Zhang, Z.-Q. *et al.* Kinetics of CD4<sup>+</sup> T cell repopulation of lymphoid tissues after treatment of HIV-1 infection. *Proc. Natl Acad. Sci. USA* **95**, 1154–1159 (1998).
57. Smith, K. Y. *et al.* Thymic size and lymphocyte restoration in HIV infected patients following 48 weeks of therapy with zidovudine, lamivudine, and zalcitabine. *J. Infect. Dis.* **181**, 141–147 (2000).
58. Kourtis, A. P. *et al.* Early progression of disease in HIV-infected infants with thymus dysfunction. *N. Engl. J. Med.* **335**, 1431–1436 (1996).
59. Auger, I. *et al.* Incubation periods for paediatric AIDS patients. *Nature* **336**, 575–577 (1988).
60. Belanger, E., Meyer, L., Carre, N., Coutellier, A. & Deveau, C. Influence of age at infection on human immunodeficiency virus disease progression to different clinical endpoints: the SEROCO cohort (1988–1994). *Int. J. Epidemiol.* **26**, 1340–1345 (1997).
61. Globerson, A. & Effros, R. B. Ageing of lymphocytes and lymphocytes in the aged. *Immunol. Today* **21**, 515–521 (2000).
62. McCune, J. M. Thymic function in HIV-1 disease. *Semin. Immunol.* **9**, 397–404 (1997).
63. McCune, J. M. *et al.* High prevalence of thymic tissue in adults with HIV-1 infection. *J. Clin. Invest.* **101**, 2301–2308 (1998).
64. Allen, T. M. *et al.* Tat-specific cytotoxic T lymphocytes select for SIV escape variants during resolution of primary viraemia. *Nature* **407**, 386–390 (2000).
65. Crowe, S. Role of macrophages in the pathogenesis of human immunodeficiency virus (HIV) infection. *Aust. NZ J. Med.* **25**, 777–783 (1995).
66. Geijtenbeek, T. B. *et al.* DC-SIGN, a dendritic cell-specific HIV-1-binding protein that enhances trans-infection of T cells. *Cell* **100**, 587–597 (2000).
67. Schluns, K. S., Kieper, W. C., Jameson, S. C. & Lefrançois, L. Interleukin-7 mediates the homeostasis of naive and memory CD8 T cells in vivo. *Nature Immunol.* **1**, 426–432 (2000).
68. Napolitano, L. A. *et al.* Increased production of IL-7 accompanies HIV-1-mediated T-cell depletion: implications for T-cell homeostasis. *Nature Med.* **7**, 73–79 (2001).
69. Fry, T. J. *et al.* A potential role for IL-7 in T cell homeostasis. *Blood* (in the press).
70. Chene, L. *et al.* Thymocyte-thymic epithelial cell interaction leads to high-level replication of human immunodeficiency virus exclusively in mature CD4<sup>+</sup> CD8<sup>+</sup> CD3<sup>+</sup> thymocytes: a critical role for tumor necrosis factor and interleukin-7. *J. Virol.* **73**, 7533–7542 (1999).
71. Deeks, S. G., Barbour, J., Martin, J., Swanson, M. & Grant, R. M. Sustained CD4 T cell response after virologic failure of protease inhibitor based regimens in HIV infected patients. *J. Infect. Dis.* **181**, 946–953 (2000).
72. Kaufmann, D., Pantaleo, G., Sudre, P. & Telenti, A. CD4-cell count in HIV-1-infected individuals remaining viraemic with highly active antiretroviral therapy. *Lancet* **351**, 732–734 (1998).
73. Deeks, S. G. *et al.* Incomplete and complete suppression of HIV-1 replication with protease inhibitor-based regimen are associated with similar T cell dynamics and activation. *Proc. Natl Acad. Sci. USA* (submitted).
74. Stoddart, C. A. *et al.* Impaired replication of protease inhibitor-resistant HIV-1 in human thymus. *Nature Med.* (in the press).
75. Casella, C. R. & Finkel, T. H. Mechanisms of lymphocyte killing by HIV. *Curr. Opin. Hematol.* **4**, 24–31 (1997).
76. Gandhi, R. T. *et al.* HIV-1 directly kills CD4<sup>+</sup> T cells by a Fas-independent mechanism. *J. Exp. Med.* **187**, 1113–1122 (1998).
77. Ahsan, N. & Langhoffer, E. Immunopathogenesis of human immunodeficiency virus. *Semin. Nephrol.* **18**, 422–435 (1998).
78. Cao, J., Park, I. W., Cooper, A. & Sodroski, J. Molecular determinants of acute single-cell lysis by human immunodeficiency virus type 1. *J. Virol.* **70**, 1340–1354 (1996).
79. Stewart, S. A., Poon, B., Jowett, J. B. M. & Chen, I. S. Y. Human immunodeficiency virus type 1 Vpr induces apoptosis following cell cycle arrest. *J. Virol.* **71**, 5579–5592 (1997).
80. Ayyavoo, V. *et al.* HIV-1 Vpr suppresses immune activation and apoptosis through regulation of nuclear factor  $\kappa$ B. *Nature Med.* **3**, 1117–1123 (1997).
81. Hellerstein, M. K. Measurement of T-cell kinetics: recent methodologic advances. *Immunol. Today* **20**, 438–441 (1999).

#### Acknowledgements

I thank J. Harris, M. Hellerstein, K. Komanduri, J. Moore, L. Napolitano, D. Nixon, K. K. Smith-McCune and Z. Grossman for helpful discussions and for their critical reading of the manuscript. I also apologize for the fact that many important references could not be included here because of space constraints. Work in my laboratory has been supported by grants from the NIH, the J. David Gladstone Institutes, the Elizabeth Glaser Pediatric AIDS Foundation and the Burroughs Wellcome Fund Clinical Scientist Award.



# Cellular immune responses to HIV

Andrew J. McMichael & Sarah L. Rowland-Jones

MRC Human Immunology Unit, Institute of Molecular Medicine, John Radcliffe Hospital, Oxford OX3 9DS, UK

**The cellular immune response to the human immunodeficiency virus, mediated by T lymphocytes, seems strong but fails to control the infection completely. In most virus infections, T cells either eliminate the virus or suppress it indefinitely as a harmless, persisting infection. But the human immunodeficiency virus undermines this control by infecting key immune cells, thereby impairing the response of both the infected CD4<sup>+</sup> T cells and the uninfected CD8<sup>+</sup> T cells. The failure of the latter to function efficiently facilitates the escape of virus from immune control and the collapse of the whole immune system.**

**T**he human immunodeficiency virus (HIV) stimulates strong immune responses by cytotoxic T lymphocytes (CTLs) in infected people<sup>1,2</sup>, despite causing profound immunodeficiency. In the acute phase of the infection, the CTL response initially follows the rise in HIV in the blood and when that response reaches a peak the virus level falls (Fig. 1); after that there is an inverse relationship between CTL response and virus load<sup>3</sup>. In the monkey model of HIV infection, CTLs can be depleted *in vivo* by infusion of antibody specific for the CD8 glycoprotein, which is a characteristic of CTLs. When these animals are infected with simian immunodeficiency virus (SIV), the early control of the virus fails<sup>4</sup>; if the antibody is instead given during the chronic phase of infection, the virus level rises until the effects of the antibody wear off<sup>5</sup>. These data imply that CTLs are important in control of the virus, a view supported by *in vitro* experiments<sup>6</sup> and by the frequent selection of virus mutants *in vivo* that are no longer recognized by CTLs<sup>7,8</sup>, and that therefore escape immune control.

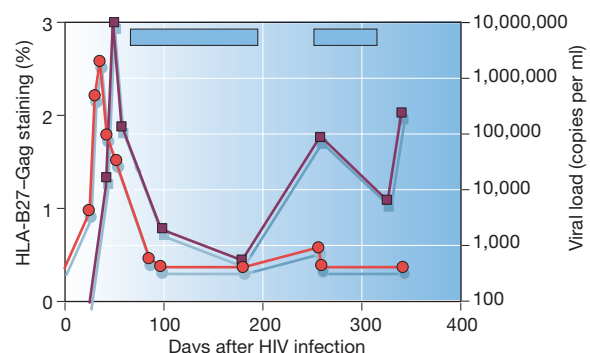
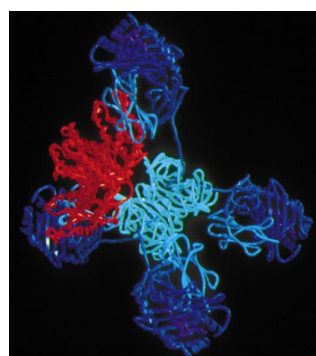
The central role of CTLs in controlling the virus is also emphasized by the influence of human leukocyte antigen (HLA) type on the rate of progress on HIV infection towards AIDS<sup>9,10</sup>. CTLs recognize virus peptides presented by HLA class I molecules, and different HLA types present different peptides and thus affect the quality of the immune

response. HLA types associated with slow progression of the infection, such as HLA-B27 and HLA-B57 (ref. 10), could stimulate more effective immune responses compared with those that confer increased susceptibility, such as HLA-B35 (ref. 9). Similarly, homozygosity at the HLA class I loci, which is also associated with more rapid progression of HIV, offers less opportunity for a diverse T-cell response<sup>9</sup>.

Although these findings indicate that the CTL response goes some way towards controlling HIV infection, they raise the question of why this response fails ultimately to deal with this virus. This is not the case for other common, persisting human virus infections, such as Epstein–Barr virus (EBV) and cytomegalovirus (CMV)<sup>11</sup>. So why is HIV different?

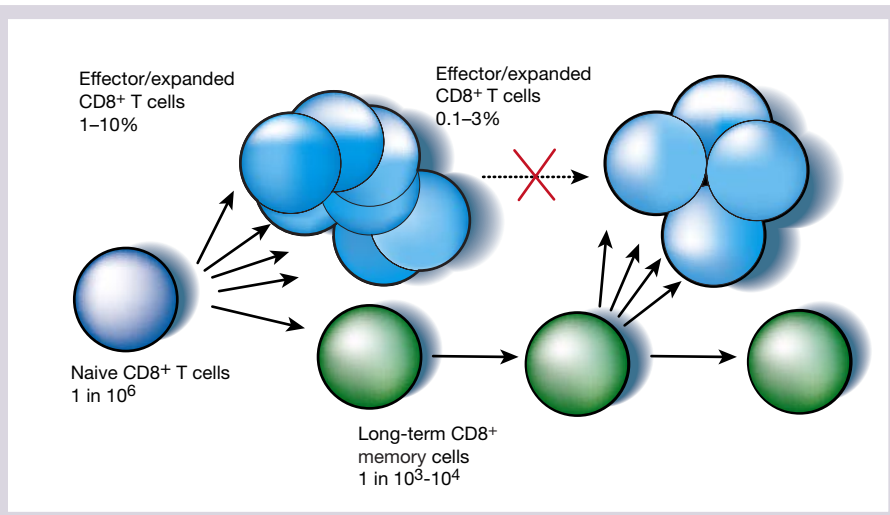
## CD8<sup>+</sup> T-cell response in acute HIV infection

Early analyses of the cellular immune response to HIV in acute infection were measured by a limiting dilution technique where T lymphocytes that responded to HIV were cloned and counted; this showed a strong CTL response coinciding with the initial peak of viraemia<sup>12,13</sup> and preceding production of any neutralizing antibody. Better quantification of the early T-cell response can be made with HLA tetramers, where four purified HLA molecules folded around an antigenic virus-derived peptide (epitope) are linked to streptavidin (Fig. 1). Such multimeric reagents bind with sufficient avidity to T-cell receptors to make it



**Figure 1** HLA tetramers can be used to quantify early T-cell responses. **a**, An HLA tetramer, based on the known structures of HLA class I (blue), streptavidin (turquoise) and phycoerythrin (red). This can be used to stain antigen-specific CD8<sup>+</sup> T cells using fluorescence-activated cell sorting. **b**, The early CTL response to HIV. The viraemia rose to a peak (red line) and was followed by the CD8<sup>+</sup> T-cell response to the immunodominant Gag 263–272 epitope presented by HLA-B27, demonstrated by the HLA-B27–epitope tetramer (mauve line). The viraemia started to fall as the CD8<sup>+</sup> T-cell response reached a peak. Further observations were influenced by the initiation of antiretroviral drug therapy (blue bands), but the eventual level of CD8<sup>+</sup> T cells specific for a single dominant epitope is typical. (Measurements from ref. 14 in collaboration with C. Workman, Sydney.)

**Figure 2** Expansion of HIV-specific CD8<sup>+</sup> T cells in acute and chronic infection. Rare antigen-specific naive CD8<sup>+</sup> T cells (blue) are stimulated to divide rapidly to generate T cells that can be detected with epitope-specific HLA-peptide tetramers (turquoise and green). The frequency of effector cells (turquoise) is much higher than those that can divide *in vitro* (green). Persisting HIV antigen maintains the expanded T cells at a high level, but in the absence of further antigen these cells tend to die by apoptosis (turquoise). The population of T cells that is capable of dividing further (green) maintains long-term memory and is likely to continually generate the expanded effectors (turquoise). The model simplifies the different routes of differentiation and there are almost certainly intermediate phenotypes (see Fig. 3).



possible to identify and count antigen-specific T cells. Using tetramers presenting an immunodominant peptide epitope, the number of HIV-specific T cells was shown to peak just after level of viraemia starts to fall<sup>14</sup>, suggesting that CTLs are responsible (Fig. 1). A similar pattern is seen in SIV-infected macaques<sup>4</sup>.

The acute T-cell response is large: as many as 10% of the T cells that carry the CD8 surface protein (which are about half of all T lymphocytes) are involved (refs 14, 15, and C. Spina and P. Hansasuta, unpublished data). Given that the frequency of naive T cells specific for any single HIV epitope must be less than one in a million, such a large reaction involves an expansion in cell numbers requiring at least 17 cell divisions over 2–3 weeks, assuming all progeny survive. In other virus infections, the acute CD8<sup>+</sup> T-cell response can be even larger<sup>16</sup>. Little is known about the function of these early HIV-specific T cells. EBV-specific CD8<sup>+</sup> T cells responding to acute infection release cytokines that kill target cells, but they also die by apoptosis or programmed cell death<sup>17</sup>, which means that their number declines rapidly if antigen does not persist<sup>18</sup>. Antigen-activated cell death probably controls CTL numbers, but there is a danger that the whole T-cell response might become terminally differentiated and be deleted. This has been seen in mice infected with aggressive lymphocyte choriomeningitis virus (LCMV)<sup>19</sup> and may occur occasionally in HIV infection<sup>20</sup>. The CD8<sup>+</sup> T cells responding in acute HIV infection are restricted to a few clones<sup>20</sup>. Similar restrictions are seen in acute EBV infection (glandular fever)<sup>16</sup>, where a few dominant clones persist into long-term memory<sup>17</sup>. In some acutely infected HIV patients, the responding CTLs are monoclonal<sup>20</sup>, but this carries a bad prognosis, possibly because it risks over-expansion and exhaustion or because it makes it easier for the virus to escape by mutation of the single epitope<sup>21</sup>.

### CD8<sup>+</sup> T-cell response in chronic HIV infection

In chronic HIV infection, the expanded HIV-specific T cells persist at high frequencies; often 1–2% of all circulating CD8<sup>+</sup> T cells are specific for a dominant HIV epitope<sup>3,22</sup>. There are similar numbers in lymph nodes<sup>23</sup> (V. Appay, unpublished data; G. Pantaleo, unpublished data), indicating that 10<sup>9</sup> CD8<sup>+</sup> T cells can be specific for a single epitope; similar T-cell responses are made to persistent EBV and CMV. These T cells probably turn over continuously<sup>24,25</sup>; and like the acutely expanded T cells, they tend to die by apoptosis *ex vivo*<sup>26</sup>. The high number of responding T cells is almost certainly dependent on continued antigen stimulation, because reduction of HIV level by combinations of potent antiretroviral drugs causes a steady decline in tetramer-stained CD8<sup>+</sup> T cells<sup>27–29</sup>. Without treatment, the high number of HIV-specific CD8<sup>+</sup> T cells often persists into late infection and they can still be detectable when AIDS develops.

As in acute infection, the expanded T cells are often oligoclonal<sup>30</sup>, distorting the distribution of T cells with different families of T-cell receptor in the blood. Persistence of T-cell clones marked by their receptors for over five years indicates that either HIV-specific CD8<sup>+</sup> T cells are long lived<sup>31,32</sup> or they are hard to exhaust if they are continuously in cycle<sup>24</sup>.

### Dynamics of the CD8<sup>+</sup> T-cell response

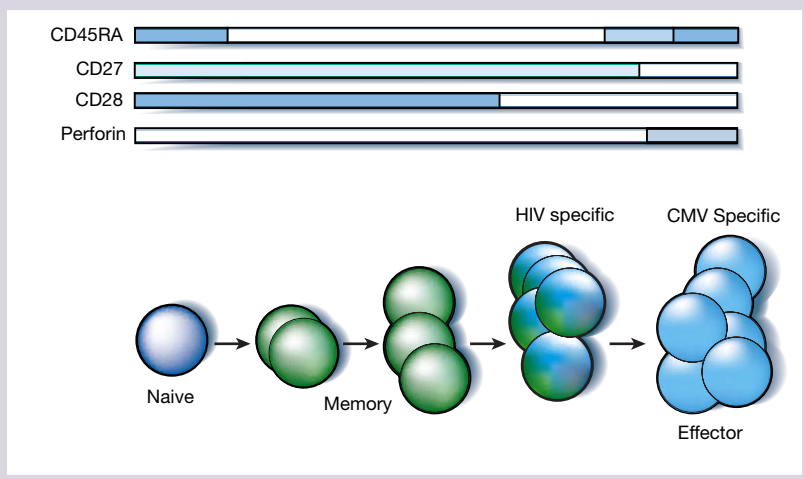
The huge number of CD8<sup>+</sup> T cells observed in acute virus infections<sup>33,34</sup> implies massive initial expansion of virus-specific CD8<sup>+</sup> T cells, which is maintained if virus persists but declines rapidly if virus is cleared<sup>15</sup>. There is a discrepancy between the number of CD8<sup>+</sup> T cells detected by tetramers and the number seen by the more traditional limiting dilution assay (LDA)<sup>3,22,35–37</sup>. The LDA requires that the cloned T cells divide at least 12 times *in vitro* to become detectable in a functional analysis, such as lysis of HIV-infected target cells. Many or most of the tetramer-stained T cells die rapidly *ex vivo*, and in most<sup>38,39</sup> but not all<sup>40</sup> studies, only about 10% can be cloned. Thus antigen-stimulated CD8<sup>+</sup> T cells, even within the same clone, can be divided into two types — terminally differentiated cells (in large numbers but likely to die) and long-term memory cells (in low numbers but able to grow). There may well be a continuum of T cells in between these extremes, in various states of differentiation (Fig. 2).

### Function of HIV-specific CD8<sup>+</sup> T cells

Virus-specific CTLs possess a range of antiviral activities, which vary in importance in different infections. These include the ability to kill infected cells and to produce cytokines and chemokines. Perforin is a protein, made by CD8<sup>+</sup> T cells, that is present in granules, and together with the granzymes is important in triggering target-cell death. Perforin-knockout mice do not recover from LCMV infection, which implies that lysis of infected cells is crucial for the control of this infection<sup>41</sup>. Although CD8<sup>+</sup> T cells are critical, these knockout mice handle other viruses effectively, such as vaccinia and hepatitis B viruses, implying that cytokine production is more important for control of these pathogens<sup>42–44</sup>.

It is not yet clear which functions of CTLs are most important in controlling HIV. They produce cytokines that can affect viral replication<sup>45,46</sup>. These include interferon- $\gamma$  (IFN- $\gamma$ ), which inhibits HIV replication<sup>47,48</sup>, and tumour-necrosis factor- $\alpha$  (TNF- $\alpha$ ), which can upregulate viral replication<sup>49,50</sup> through activation of the HIV promoter in the virus 5' long terminal repeat (LTR). HIV-specific CTLs also produce the CC chemokines MIP-1 $\alpha$ , MIP-1 $\beta$  and RANTES<sup>51,52</sup>, which suppress HIV replication<sup>53</sup> by competition for, or downregulation of, CCR5 — the cellular co-receptor (with CD4) for the virus. HIV-specific CTLs secrete these antiviral factors at sites of virus replication, efficiently inhibiting virus replication *in vitro*<sup>6</sup>.

**Figure 3** Differentiation of CD8<sup>+</sup> T cells. Surface glycoproteins CD45 (RA and RO isotypes), CD27 and CD28 distinguish stages of CD8<sup>+</sup> T-cell maturation. Perforin is expressed only at the final stage. CD45RA reappears as these T cells mature to full effectors. The scheme represents a broad pattern for most HIV- and CMV-specific T cells, although there are some exceptions.



Inhibition involves lytic mechanisms and release of CC chemokines and a partially characterized secreted factor (CD8<sup>+</sup> T-cell antiviral factor, or CAF)<sup>54,55</sup>. CAF blocks LTR-mediated transcription in infected cells<sup>56</sup>, thereby shutting off virus production. It may, in addition, facilitate the development of latency, where HIV complementary DNA is integrated into the host-cell genome but remains silent and expresses no messenger RNA and no protein.

Cultured HIV-specific CTLs have been shown to lyse HIV-infected CD4<sup>+</sup> T cells *in vitro*<sup>57</sup>, despite the ability of the virus Nef protein to reduce expression of class I HLA molecules on their surface<sup>58</sup>. Susceptibility of infected cells to lysis by CTL clones parallels expression of viral proteins such as Gag p24 within the cell and precedes production of extracellular virus<sup>59</sup>. This means that lysis is a potent weapon against HIV. Lysis is mediated predominantly by perforin and granzymes<sup>60</sup>, although a minority of CTLs use the alternative route where Fas ligand expressed on antigen-activated CTLs triggers apoptosis in the target cells that also express Fas<sup>61</sup>.

Findings from several studies show that virus-specific CTLs taken *ex vivo* can have functional defects that could undermine their control of virus<sup>62,63</sup>. Although tetramer staining indicates large numbers of HIV-specific T cells are present in acute or chronic HIV infection, this technique alone makes no measurement of their function. When HIV epitope/HLA class I tetramer-staining was combined with intracellular staining for cytokines and chemokines, it was found that most HIV-specific cells in patients with chronic HIV disease produced IFN- $\gamma$ , TNF- $\alpha$  and MIP-1 $\beta$  on contact with their cognate antigen *ex vivo*<sup>64</sup>. This pattern of cytokine secretion by HIV-specific cells was similar to that of CMV-specific cells from HIV-uninfected donors. However, a striking difference was seen in the level of intracellular perforin. Less than 15% of HIV-specific cells contained perforin, which was reflected in poor *ex vivo* killing of appropriate target cells, compared with CMV-specific cells from the same donors, 50% of which expressed perforin and killed well<sup>64</sup>. An earlier study had shown low levels of perforin in CD8<sup>+</sup> T cells in the lymph nodes of infected patients<sup>65,66</sup>.

Thus, HIV-specific T cells may be less efficient killer cells than expected, a feature that might not necessarily have been detected in experiments with cultured CTL clones, cited above, because culture can readily modify or select for function. It is uncertain why HIV-specific cells poorly express perforin: these cells lack expression of the glycoprotein CD28 on their surface, but they retain CD27 (ref. 64). In contrast, CMV-specific cells lose expression of both molecules and this loss is thought to mark out mature effector cells<sup>67</sup>. So, HIV-specific CD8<sup>+</sup> T cells *in vivo* may be immature<sup>67</sup> rather than end-stage effectors as first thought (Fig. 3). Recent findings exploring expression of two other surface glycoproteins — CCR7, a lymphoid-organ homing receptor, and CD45RA, a long isotype of CD45 — support this view: immature CD8<sup>+</sup> T memory cells were CCR7<sup>+</sup>CD45RA<sup>+</sup>,

fully mature CD8<sup>+</sup> T cells specific for CMV were CCR7<sup>+</sup>CD45RA<sup>+</sup>, and HIV-specific CD8<sup>+</sup> T cells were CCR7<sup>+</sup>CD45RA<sup>+</sup> (ref. 68). Failure of CD8<sup>+</sup> T cells to mature could be a consequence of impaired T-cell help, as suggested by experiments in mice infected with LCMV and lacking helper T cells, where responding T cells were also deficient in function (see below)<sup>62</sup>, although the nature or strength of the antigenic stimulus could also be relevant. It will be interesting to determine whether different CTL functional phenotypes are associated with different clinical outcomes, and whether such diversity might be linked to the presenting HLA type, particularly those associated with good or bad clinical outcomes<sup>9,10</sup>.

### Do HIV-specific CTLs fail for lack of help?

These results question the role of the HIV-specific helper T-cell response, mediated by T cells that express CD4 rather than CD8. The question arises because of the susceptibility of the former to HIV infection. Poor CD4<sup>+</sup> T-cell responses to HIV and other recall antigens, measured by proliferation of the T cells to antigen *in vitro*, were found in HIV-infected patients several years ago<sup>69</sup>. But a minority of HIV-infected people, who maintain high CD4<sup>+</sup> T-cell counts in their blood for many years, actually made good proliferative responses to Gag p24 (ref. 70). Initiation of highly active antiretroviral therapy (HAART) early in infection rescues an anti-Gag p24 CD4<sup>+</sup> T-cell proliferative response in most patients<sup>70</sup>. This result is consistent with the observation that HIV-specific CD4<sup>+</sup> T cells, detected by rapid antigen-stimulated release of IFN- $\gamma$ , were found in early infection but disappeared in untreated patients<sup>71</sup>. However, Pitcher *et al.*<sup>72</sup> used a different technique to detect CD4<sup>+</sup> T cells responding to Gag p24 by IFN- $\gamma$  secretion in patients at all stages of HIV infection, although the HIV-specific T cells were much less frequent than CMV-specific CD4<sup>+</sup> T cells in the same people, so there is agreement that the number of responding T cells is low.

CD4<sup>+</sup> T cells that are specific for HIV may be exceptionally susceptible to attack and destruction by HIV. In the periphery, HIV binds to DC-SIGN, a glycoprotein expressed on dendritic cells, without infecting them<sup>73</sup>. This enables dendritic cells to bring HIV into draining lymph nodes early in primary infection. There the dendritic cells form foci, surrounded by CD4<sup>+</sup> T cells activated by antigen processed and presented by dendritic cells in their HLA class II molecules. The proximity of HIV-associated or HIV-infected dendritic cells with activated CD4<sup>+</sup> T cells, which are especially susceptible to infection, is a lethal mix for the latter<sup>74</sup>. HIV-specific T cells are thus attacked by HIV and are likely to be specifically deleted early in infection. There is evidence that this process can be reversed when acutely infected patients are given HAART<sup>70,75</sup>; this results in preservation of proliferative CD4<sup>+</sup> T cells and ultimately better control of HIV on or off drug treatment.

The extent to which the impaired CD4<sup>+</sup> T-cell help influences the CD8<sup>+</sup> T-cell response is now becoming apparent (Fig. 4). Early in



untreated HIV infection, some CD4<sup>+</sup> T-cell help may be present and the initial CD8<sup>+</sup> T-cell response may be similar to that against any other virus. The problems may come later. T-cell help is known to be important for priming the CD8<sup>+</sup> T-cell responses<sup>76</sup>, for maintaining CD8<sup>+</sup> T-cell memory<sup>77</sup> and for maturing CD8<sup>+</sup> T-cell function<sup>62</sup>.

Evidence that CD4<sup>+</sup> T cells are important in priming CD8<sup>+</sup> T cells comes from studies in mice that pinpoint CD40L, expressed on activated CD4<sup>+</sup> T cells, as crucial in triggering dendritic cells to produce the cytokine interleukin (IL)-12, which in turn is central in initiating the CD8<sup>+</sup> T-cell response<sup>76</sup>. Some viruses can by-pass this step by activating dendritic cells directly<sup>76</sup>, but it is not known whether HIV can do this. Failure of T-cell help could disrupt the ability of CD8<sup>+</sup> T cells to make new primary immune responses to mutant viruses that have evaded the originally dominant CTL responses (see below).

Helper T cells may be important for the maintenance of CD8<sup>+</sup> memory. Survival of infused CD8<sup>+</sup> T-cell clones in CMV-infected patients depends on the presence of specific CD4<sup>+</sup> T-cell help<sup>77</sup>; adoptively transferred HIV-specific CTLs survive poorly in HIV-infected patients<sup>78,79</sup>. These T-cell clones used in adoptive transfer studies may be particularly dependent on IL-2 and other factors provided by helper T cells. In contrast, the long survival of particular CD8<sup>+</sup> T-cell clones for many years in patients chronically infected with HIV<sup>32</sup>, when CD4<sup>+</sup> T-cell help is known to be damaged, suggests that naturally activated CD8<sup>+</sup> T cells may be able to survive better in the absence of help.

The third role for CD4<sup>+</sup> T cells in regulating CD8<sup>+</sup> T-cell function has been shown in animal models where absence of CD4<sup>+</sup> cells resulted in a CD8<sup>+</sup> T-cell response that was numerous in terms of HLA tetramer-stained cells, but lacking in function<sup>62</sup>. This seems particularly pertinent to HIV infection in view of the findings, discussed above, that CD8<sup>+</sup> T cells in HIV infection are not fully mature<sup>64</sup>.

It is clear that patients with high CD4<sup>+</sup> T-cell numbers and detectable helper T cells do better, which is consistent with a more effective T-cell immune response and better immune control of the HIV, demonstrated by a low virus load. Current attempts to rescue the helper T-cell response with drug treatment and vaccine immunotherapy may enhance the CD8<sup>+</sup> T-cell response with real benefits<sup>75</sup>. It should also be recognized that CD4<sup>+</sup> T cells could have direct antiviral effects, releasing antiviral cytokines and chemokines and killing infected cells. Activated human T cells express class II HLA, so HIV-infected cells should be targets for CD4<sup>+</sup> T cells, although the HIV protein Tat can downregulate HLA class II at the surface of HIV-infected cells<sup>80</sup>.

The role of helper T cells in CTL survival and function is an area that needs more attention. It is certain that CD4<sup>+</sup> T-cell function is impaired early in infection, even if the degree of dysfunction is in dispute. Later, CD4<sup>+</sup> T-cell numbers decline drastically. Failure of T-cell help must be central in the pathogenesis of HIV infection and could be the critical feature that ultimately undermines immune control. It will be important to determine when T-cell help is lost in relation to the early events in HIV infection that set the virus load and prognosis.

### Escape of HIV from the cellular immune response

CD8<sup>+</sup> T cells can act against HIV most effectively by killing infected cells before they generate new virus particles. There is a window of about 12 hours from the onset of virus protein synthesis to the budding of new virus particles. During this period, epitope peptides are presented by HLA class I molecules at the surface of infected cells and can be recognized by CTLs and lysed within five hours<sup>77</sup>. Unless the killing process eliminates HIV rapidly, it should exert a selective force, giving an advantage to cells infected with viruses that have mutated critical amino acids in the dominant epitopes. These infected cells escape lysis and propagate the mutant virus. Selection of such mutants could be especially favoured when there is impairment of CTL function<sup>64</sup>, so that virus level is higher than it might otherwise be, with more virus replication and therefore more mutation. The availability of potential escape mutants is the product of virus

**Table 1 Well characterized CTL escape mutations in HIV-infected patients**

| HLA | HIV gene       | Sequence                                                             | Effect                   | Reference |
|-----|----------------|----------------------------------------------------------------------|--------------------------|-----------|
| A3  | <i>nef</i>     | QVPLRPMTYK<br>.....                                                  | Deleted                  | 77        |
| B8  | <i>nef</i>     | FLKEKGGL<br>FLKE <b>X</b> GGL<br>.....                               | No binding<br>Deleted    | 78        |
| B44 | <i>env</i>     | AENLWVTY<br>A <b>X</b> NLWVTY                                        | No binding               | 79        |
| B27 | <i>gag</i> p24 | KRWIILGLNK<br>K <b>K</b> WIIMGLNK (× 4)<br>K <b>G</b> WIILGLNK (× 1) | No binding<br>No binding | 6,80,82   |

The original sequence of the epitope is shown in the standard single-letter amino-acid code. The mutations observed are shown underneath with the changed amino acid in bold. An 'X' indicates two or more variant amino acids; dotted lines indicate that the region of the epitope was deleted; (× 4) and (× 1) mean that the mutation was found in four different people and one person, respectively. In the HLA-A3 *nef* example the *nef*-deleted mutant was found in 30% of patient-derived HIV sequences after immunotherapy with a T-cell clone specific for that epitope. In the other cases the mutants completely replaced the original sequence on two or more consecutive occasions.

replication and mutation rate, so the balance between the strength of killing and the level of virus replication should determine how frequently virus escape mutants arise (Fig. 5). It has been argued that the inverse relationship between CTL number (detected by tetramers) and virus load implies the existence of defective killers, because at equilibrium a positive correlation would be expected, such as that found for T cells specific for the non-immunosuppressive human T-cell leukaemia virus type 1 (ref. 81). The observed and expected findings can be resolved if HIV starts to suppress the CTL response above a certain virus threshold. CTLs would still suppress virus load but would be progressively less efficient at doing so as virus levels rise (discussed in ref. 82).

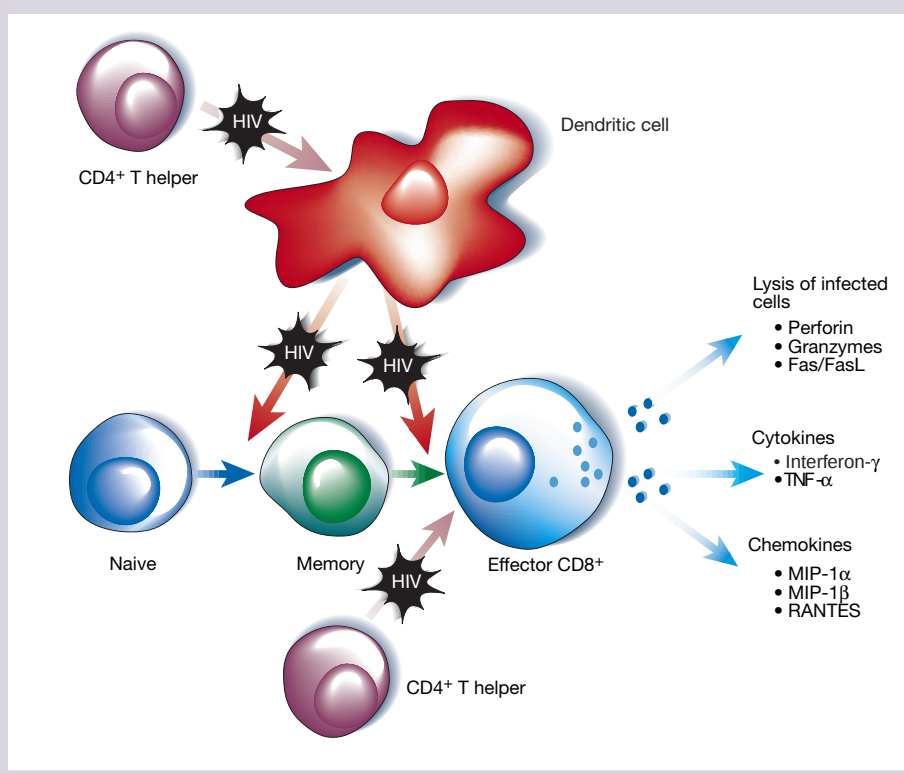
Selection of mutants by CTLs is probably one of the main features of HIV infection. Longitudinal studies of individual patients, matching dominant CTL responses with changes in amino-acid sequence, have identified clear cases where a single change has abrogated presentation by the class I HLA molecules and these viruses have become dominant in the quasispecies, the swarm of virus variants that arise *in vivo* (Table 1). A good example was found after an attempt to treat a patient with his own CTL clone, infused in very high numbers<sup>83</sup>. The clone was specific for a Nef epitope presented by HLA-A3, and virus was selected that had deleted the relevant region of *nef*. This mutated virus accounted for 30% of the total virus, an impressive result given that *nef*-deleted viruses are often crippled compared with wild-type viruses. Selection and fixation of escape mutants in acute HIV infection, when virus turnover is high, have been described<sup>84,85</sup>. In each case, the immunodominant epitope was altered so that it did not bind to the presenting HLA molecule.

The same escape mutation in the immunodominant Gag p24 epitope presented by HLA-B27 has been seen in more than eight different patients (ref. 86 and Goulder *et al.*, personal communication). Study of this mutation and its selection by T cells is instructive of the whole phenomenon. First, the epitope is strongly immunodominant, and in patients with HLA-B27, T cells specific for this antigen are strongly selected. Three of four escape mutations described by Kelleher *et al.*<sup>87</sup> occurred in late infection as virus levels rose, and one occurred in acute infection<sup>87</sup>. The commonest change was an arginine-to-lysine substitution that abrogates binding to the

**Table 2 Mechanisms by which HIV evades the CD8<sup>+</sup> T-cell immune response**

|                                |                                       |
|--------------------------------|---------------------------------------|
| Provirus latency               | No antigen expression                 |
| Sequestration                  | Invisible antigen                     |
| Switch to CXCR4 receptor       | Insensitive to CC chemokines          |
| Downregulation of HLA class I  | Invisible to CD8 <sup>+</sup> T cells |
| Downregulation of HLA class II | Invisible to CD4 <sup>+</sup> T cells |
| Upregulation of Fas ligand     | Back-killing of T cells               |
| Epitope mutation               | Non-recognition                       |
|                                | Antagonism of T cells                 |

**Figure 4** CD4<sup>+</sup> T-cell dependence of CD8<sup>+</sup> T cells. CD4<sup>+</sup> T cells are important for priming dendritic cells to initiate CD8<sup>+</sup> T-cell responses. They help maintain memory T cells and are important in maturation of CD8<sup>+</sup> T-cell function. All of these actions are impaired by HIV infection. In addition, HIV can directly infect and impair dendritic-cell function.



HLA-B27 molecule<sup>86</sup>; the lysine had not been seen before in B-clade Gag p24 sequences. However, it was found in B27<sup>+</sup> patients only when there was a second change in the epitope, a leucine-to-methionine substitution<sup>87</sup>, four residues downstream. Recently, a third amino-acid change in the same conformational region of the p24 capsid has been found to be necessary (Kelleher *et al.*, unpublished results). The mutant sites within the epitope are located on the same face of the seventh alpha-helix of the amino-terminal domain of the p24 capsid and they probably complement each other in the capsid structure. The necessity for multiple mutations could explain why these escape mutations usually occurred late in infection. This region of Gag p24 is well conserved and important in the packaging of the capsid<sup>88,90</sup>, so relatively few mutations may be compatible with viable virus<sup>90</sup>. HLA-B27 and B57, which are both associated with slow progression to AIDS, select epitopes in this region of p24. Thus, the ease of escape probably depends on the site of the epitope in the natural protein and this could account for different rates of disease progression associated with different HLA types<sup>9,10</sup>.

There is also circumstantial evidence for CTL epitope escape in cross-sectional studies where few or no sequential measurements are made (the commonest type of study). For example, Phillips *et al.*<sup>7</sup> described changes occurring during ongoing infection in more than one epitope presented by HLA-B8. Goulder *et al.*<sup>91</sup> found two HIV-infected, HLA-identical haemophiliac brothers who made CTL responses to totally different epitopes, arguing that this was a consequence of mutation in the normally immunodominant epitopes recognized by CTLs in one of the brothers. This resulted in a completely different pattern of response, offering one explanation for the complexity often seen in HIV-specific T-cell responses in chronic infection, where HIV-specific CTLs in patients of similar HLA type may respond to very different epitopes. These findings are consistent with evidence that escape from one CTL population is followed by a new CTL response to a new epitope<sup>92</sup>; this may weaken immune control over the virus because the subdominant CTL response is less effective, although this has not been shown clearly. More serious to the patient may be the eventual inability to make a new primary CTL responses when CD4<sup>+</sup> T-cell help is damaged severely.

Studies in macaques have strengthened the notion that escape from CTLs is frequent and therefore important. Macaques infected with cloned SIV showed multiple mutations in the immunodominant CTL epitopes during the first few months of infection<sup>8</sup>. Up to eight mutations in Nef and Env epitopes were selected simultaneously. There was also clear selection for non-synonymous nucleotide changes and the haplotype of the major histocompatibility complex was critical in the selection process, strongly implicating CTLs<sup>8</sup>. Later, Allen *et al.*<sup>92</sup> found that mutations at different epitopes could occur at different rates. They argued that the CTL response to the variable Tat epitope was strongly selective and therefore more protective than an equally strong response to a Gag p27 epitope that varied little. However, the Tat-specific response is of little value because the epitope alters so rapidly. These data in macaques establish escape mutation as a normal event in HIV infection, undermining control by CTLs.

These results also imply that HLA type can mould the virus. The immune system can respond to many epitopes, but a hierarchy of immunodominance exists that is exposed by the escape mutations<sup>93</sup>. Given that the predominant HLA types differ markedly in different populations, is it possible that the common HLA types in different HIV-exposed populations could have contributed to the generation of the different clades of HIV? The clades of virus have striking geographical distributions and infect different populations<sup>94</sup>, yet they all probably arose from the same source in Central Africa.

### Other escape routes

Nef causes downregulation of HLA class I molecules at the surface of HIV-infected cells<sup>58</sup> by re-routing the newly synthesized molecules to clathrin-coated pits for endosomal degradation as they leave the trans-Golgi network<sup>95</sup>. The effect is dependent on a sequence motif in the cytoplasmic tail of the classical HLA-A and HLA-B locus molecules; HLA-C and HLA-E do not have this motif and are not downregulated<sup>96</sup>. Thus HIV-infected cells may escape attack by HLA-A- and HLA-B-restricted CTLs. Natural killer (NK) cells express inhibitory receptors for HLA-C and HLA-E and so HIV-infected cells can also evade attack by NK cells. HIV-specific CTLs kill virus-infected cells that have failed to downregulate HLA class I and so select HLA-negative cells *in vitro*<sup>58</sup>. The loss of HLA from the cell

surface takes about 48 hours, which may limit the effectiveness for escape, but Nef interferes with newly synthesized and HIV-peptide-loaded HLA molecules from a much earlier time. The strong CD8<sup>+</sup> T-cell response (in terms of number of responding cells) to HIV in the blood cannot be taken as evidence that Nef-mediated HLA downregulation is unimportant, because the response *in vivo* could reflect cross-priming of CTLs by dendritic cells<sup>97</sup> that take up HIV proteins from infected cells without risking HLA downregulation by Nef. However, the susceptibility of HIV-infected cells that expresses Nef to T-cell mediated lysis indicates that downregulation is not complete and the protection from CTL attack is only partial<sup>58</sup>.

Upregulation of Fas ligand is another consequence of Nef activity in infected cells<sup>98,99</sup>. Almost all HIV-specific T cells express Fas and so could be targets for killing by the FasL pathway or some other inhibitory effect.

Other means by which HIV can escape CTL attack include sequestration of infected cells in the central nervous system where T cells normally have no access. Any cells that are latently infected will also be invisible to the CTLs. Transition of virus from the R5 to the X4 type, when the co-receptor requirement is changed from CCR5 (natural receptor for the CC chemokines) to CXCR4 (natural receptor for stromal-derived factor) during disease progression, will mean that the virus becomes insensitive to inhibition by the CC chemokines released by antigen-activated T cells. Because HIV escape by epitope mutation would be more effective when target cells are escaping CTL killing than when virus entry is inhibited by chemokine competition, the virus change to the X4 type might favour CTL escape.

Overall there are more than enough mechanisms to explain why CTLs fail to control HIV (Table 2). Many of these processes, such as decreasing HLA class I expression, are used by many viruses, but others are special to HIV, particularly the effects of impaired helper T-cell function. Escape by epitope mutation has been reported for other viruses (reviewed in ref. 100) but seems to be particularly important for HIV and SIV.

### Implications for vaccines

There is a desperate global need for a prophylactic AIDS vaccine, but it has been difficult to find candidate vaccines that stimulate effective neutralizing antibodies<sup>101</sup>. Induction of CD8<sup>+</sup> T cells might offer a chance of partial or complete protection from HIV infection. CD8<sup>+</sup> T cells cannot prevent infection of cells by HIV, but they could abort an infection before it becomes established, or contain the virus at a significantly lower level if HIV is not eliminated. A recent study of a DNA vaccine in macaques, aimed at stimulating cellular immune responses, resulted in virus levels 1,000-fold lower after SIV challenge than in unvaccinated control animals<sup>102</sup>. If the vaccine does not persist, the HIV-specific T cells may exist only as long-term memory cells, rather than as activated effectors. HIV exposure and infection would stimulate a secondary immune response that may take some days before fully active killer cells are generated. Memory CD8<sup>+</sup> T cells can respond by releasing IFN- $\gamma$  and chemokines within six hours and divide more rapidly than naive T cells<sup>103</sup>. Furthermore, their numbers are likely to be 100–1,000 times higher than HIV-specific naive T cells in a person unexposed to HIV or a vaccine<sup>37</sup>. Theoretically, therefore, a rapid response in the memory T-cell population<sup>103</sup> could terminate the infection.

In the macaque, vaccines that stimulate CD8<sup>+</sup> T-cell responses can partially protect against challenge with SIV<sup>104,105</sup>. Complete protection is rare, but reduction of virus load compared with controls after SIV challenge is seen consistently<sup>92,102</sup>. When the CTL response is large, the reduction in the level of virus at a set point after challenge is impressive and is accompanied by real survival advantage<sup>102</sup>. This is encouraging given that aggressive-challenge viruses are used at infecting doses that are 10–100 times those of a human sexual contact.

Further encouragement for vaccines that stimulate cellular immunity derives from the finding that some commercial sex workers in Nairobi are resistant to HIV infection. These stand out in a

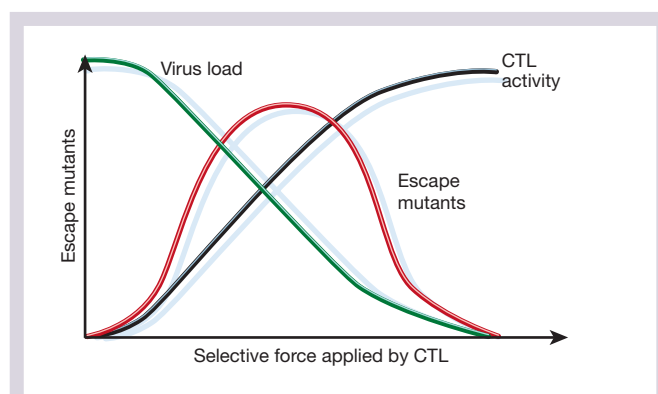
population in which 90% are infected with HIV. Most of these women make CD8<sup>+</sup> T-cell responses to HIV without making serum anti-HIV antibody<sup>106</sup>. Similar induction of CTLs without antibody has been seen in macaques that have been treated with antiretroviral drugs soon after SIV challenge<sup>107,108</sup> and these animals were protected from SIV challenge.

These findings encourage moves to start phase 1 trials in humans of vaccines that stimulate CTLs. Priming with DNA and then boosting with the same DNA construct in recombinant pox viruses stimulates strong CD8<sup>+</sup> T-cell responses in macaques<sup>18,104,109</sup>. DNA together with IL-2 was also effective at stimulating CD8<sup>+</sup> T-cell responses and protected against SIV challenge<sup>102</sup>. Serious attempts at vaccine development have to take a global perspective focusing on A- and C-clade vaccines — the clades that cause most infections worldwide. For developing countries a vaccine is the only real hope of controlling the devastating HIV pandemic that is sweeping sub-Saharan Africa and Asia.

Would CTL-inducing vaccines select escape mutants and therefore be ineffective? This is clearly possible<sup>110</sup>, but the vaccine-primed response might have advantages over natural infection that could minimize this problem by keeping virus levels low (Fig. 5). The number of antigen-specific precursors is far greater in vaccine recipients and they should respond with chemokine and cytokine expression much more rapidly<sup>103</sup>. The vaccines can also stimulate helper T cells and should therefore generate a more effective immune response which is better differentiated towards true CTLs. If a small amount of infecting virus can stimulate a strong secondary CTL response it is possible that the infection could be terminated, but this still has to be proven.

### Conclusions

The cellular immune response to HIV is complex. The CD4<sup>+</sup> T-cell response in HIV infection has long been known to be poor and we now know that it is damaged specifically early in primary HIV infection. Although earlier studies emphasized the strength of the CD8<sup>+</sup> T-cell response, more recent data have questioned whether this is optimal. Not only may the CD8<sup>+</sup> T-cell response be defective, but that impairment could facilitate selection of virus escape mutants that make it even harder for the immune system to maintain control over the virus. There is a link between the weak CD4<sup>+</sup> T-cell response and the suboptimal CD8<sup>+</sup> T-cell response to HIV. As the infection progresses, both T-cell responses decline further and immune control of the virus collapses. Vaccines that can stimulate both CD4<sup>+</sup> and CD8<sup>+</sup> T-cell responses to HIV may be able to control the virus early in infection before it causes major immune damage. □



**Figure 5** Schematic representation of the relationship between virus replication rate, manifested as virus load, the level of CTL response and the selective pressure on epitope escape mutants. When virus load is high, as in acute infection, there is no selective force until the CTLs appear. When the CTLs are maximally effective there may be little net selection because the virus load has been depressed to a low level. In chronic HIV infection, the CTLs may be suboptimal giving rise to many escape mutants because of high virus turnover.



1. Walker, B. D. *et al.* HIV-specific cytotoxic T lymphocytes in seropositive individuals. *Nature* **328**, 345–348 (1987).
2. Plata, F. *et al.* AIDS virus specific cytotoxic T lymphocytes in lung disorders. *Nature* **328**, 348–351 (1987).
3. Ogg, G. S. *et al.* Quantitation of HIV-1-specific cytotoxic T lymphocytes and plasma viral RNA load. *Science* **279**, 2103–2106 (1998).
4. Schmitz, J. E. *et al.* Control of viremia in simian immunodeficiency virus infection by CD8<sup>+</sup> lymphocytes. *Science* **283**, 857–860 (1999).
5. Jin, X. *et al.* Dramatic rise in plasma viremia after CD8(+) T cell depletion in simian immunodeficiency virus-infected macaques. *J. Exp. Med.* **189**, 991–998 (1999).
6. Yang, O. O. *et al.* Suppression of human immunodeficiency virus type 1 replication by CD8<sup>+</sup> cells: evidence for HLA class I-restricted triggering of cytolytic and noncytolytic mechanisms. *J. Virol.* **71**, 3120–3128 (1997).
7. Phillips, R. E. *et al.* Human immunodeficiency virus genetic variation that can escape cytotoxic T cell recognition. *Nature* **354**, 453–459 (1991).
8. Evans, D. T. *et al.* Virus-specific cytotoxic T-lymphocyte responses select for amino-acid variation in simian immunodeficiency virus Env and Nef. *Nature Med.* **5**, 1270–1276 (1999).
9. Carrington, M. *et al.* HLA and HIV-1: heterozygote advantage and B\*35-Cw\*04 disadvantage. *Science* **283**, 1748–1752 (1999).
10. Kaslow, R. A. *et al.* Influence of combinations of human major histocompatibility complex genes on the course of HIV-1 infection. *Nature Med.* **2**, 405–411 (1996).
11. Riddell, S. R. & Greenberg, P. D. Principles for adoptive T cell therapy of human viral diseases. *Annu. Rev. Immunol.* **13**, 545–586 (1995).
12. Koup, R. A. *et al.* Temporal association of cellular immune responses with the initial control of viremia in primary human immunodeficiency virus type 1 syndrome. *J. Virol.* **68**, 4650–4655 (1994).
13. Borrow, P., Lewicki, H., Hahn, B. H., Shaw, G. M. & Oldstone, M. B. Virus-specific CD8<sup>+</sup> cytotoxic T-lymphocyte activity associated with control of viremia in primary human immunodeficiency virus type 1 infection. *J. Virol.* **68**, 6103–6110 (1994).
14. Wilson, J. D. *et al.* Direct visualization of HIV-1-specific cytotoxic T lymphocytes during primary infection. *AIDS* **14**, 225–233 (2000).
15. Kuroda, M. J. *et al.* Emergence of CTL coincides with clearance of virus during primary simian immunodeficiency virus infection in rhesus monkeys. *J. Immunol.* **162**, 5127–5133 (1999).
16. Callan, M. F. C. *et al.* Large clonal expansions of CD8<sup>+</sup> T cells in acute infectious mononucleosis. *Nature Med.* **2**, 205–211 (1996).
17. Callan, M. F. *et al.* CD8(+) T-cell selection, function, and death in the primary immune response in vivo. *J. Clin. Invest.* **106**, 1251–1261 (2000).
18. Hanke, T. *et al.* Effective induction of simian immunodeficiency virus-specific cytotoxic T lymphocytes in macaques by using a multipeptide gene and DNA prime-modified vaccinia virus Ankara boost vaccination regimen. *J. Virol.* **73**, 7524–7532 (1999).
19. Moskopidhis, D., Lechner, F., Pircher, H. & Zinkernagel, R. M. Virus persistence in acutely infected immunocompetent mice by exhaustion of antiviral cytotoxic effector T cells. *Nature* **362**, 758–761 (1993).
20. Pantaleo, G. *et al.* Major expansion of CD8<sup>+</sup> T cells with a predominant Vb usage during the primary immune response to HIV. *Nature* **370**, 463–467 (1994).
21. Pantaleo, G. *et al.* The qualitative nature of the primary immune response to HIV infection is a prognosticator of disease progression independent of the initial level of plasma viremia. *Proc. Natl Acad. Sci. USA* **94**, 254–258 (1997).
22. Altman, J. *et al.* Direct visualization and phenotypic analysis of virus-specific T lymphocytes in HIV-infected individuals. *Science* **274**, 94–96 (1996).
23. Kuroda, M. J. *et al.* Comparative analysis of cytotoxic T lymphocytes in lymph nodes and peripheral blood of simian immunodeficiency virus-infected rhesus monkeys. *J. Virol.* **73**, 1573–1579 (1999).
24. Meynard, L. *et al.* Programmed death of T cells in HIV-1 infection. *Science* **257**, 217–219 (1992).
25. Hazenberg, M. D. *et al.* Increased cell division but not thymic dysfunction rapidly affects the T-cell receptor excision circle content of the naive T cell population in HIV-1 infection. *Nature Med.* **6**, 1036–1042 (2000).
26. Lewis, D. E., Tang, D. S., Adu-Oppong, A., Schober, W. & Rodgers, J. R. Energy and apoptosis in CD8<sup>+</sup> T cells from HIV-infected persons. *J. Immunol.* **153**, 412–420 (1994).
27. Gray, C. M. *et al.* Frequency of class I HLA-restricted anti-HIV CD8<sup>+</sup> T cells in individuals receiving highly active antiretroviral therapy (HAART). *J. Immunol.* **162**, 1780–1788 (1999).
28. Ogg, G. S. *et al.* Decay kinetics of human immunodeficiency virus-specific effector cytotoxic T lymphocytes after combination antiretroviral therapy. *J. Virol.* **73**, 797–800 (1999).
29. Kalams, S. A. *et al.* Levels of human immunodeficiency virus type 1-specific cytotoxic T-lymphocyte effector and memory responses decline after suppression of viremia with highly active antiretroviral therapy. *J. Virol.* **73**, 6721–6728 (1999).
30. Wilson, J. D. *et al.* Oligoclonal expansions of CD8(+) T cells in chronic HIV infection are antigen specific. *J. Exp. Med.* **188**, 785–790 (1998).
31. Kalams, S. A. *et al.* T cell receptor usage and fine specificity of human immunodeficiency virus 1-specific cytotoxic T lymphocyte clones: analysis of quasispecies recognition reveals a dominant response directed against a minor in vivo variant. *J. Exp. Med.* **183**, 1669–1679 (1996).
32. Moss, P. A. H. *et al.* Persistent high frequency of human immunodeficiency virus-specific cytotoxic T cells in peripheral blood of infected donors. *Proc. Natl Acad. Sci. USA* **92**, 5773–5777 (1995).
33. Murali-Krishna, K. *et al.* Counting antigen-specific CD8 T cells: a reevaluation of bystander activation during viral infection. *Immunity* **8**, 177–187 (1998).
34. Butz, E. A. & Bevan, M. J. Massive expansion of antigen-specific CD8<sup>+</sup> T cells during an acute virus infection. *Immunity* **8**, 167–175 (1998).
35. Gotch, F. M., Nixon, D. F., Alp, N., McMichael, A. J. & Borysiewicz, L. K. High frequency of memory and effector gag specific cytotoxic T lymphocytes in HIV seropositive individuals. *Int. Immunol.* **2**, 707–712 (1990).
36. Tan, L. C. *et al.* A re-evaluation of the frequency of CD8<sup>+</sup> T cells specific for EBV in healthy virus carriers. *J. Immunol.* **162**, 1827–1835 (1999).
37. Carmichael, A., Jin, X., Sissons, P. & Borysiewicz, L. Quantitative analysis of the human immunodeficiency virus type 1 (HIV-1)-specific cytotoxic T lymphocyte (CTL) response at different stages of HIV-1 infection: differential CTL responses to HIV-1 and Epstein-Barr virus in late disease. *J. Exp. Med.* **177**, 249–256 (1993).
38. Dunbar, P. R. *et al.* Direct isolation, phenotyping and cloning of low-frequency antigen-specific cytotoxic T lymphocytes from peripheral blood. *Curr. Biol.* **8**, 413–416 (1998).
39. Pantaleo, G., Koenig, S., Baseler, M., Lane, H. C. & Fauci, A. S. Defective clonogenic potential of CD8<sup>+</sup> T lymphocytes in patients with AIDS. Expansion in vivo of a nonclonogenic CD3<sup>+</sup>CD8<sup>+</sup>DR<sup>+</sup>CD25<sup>+</sup> T cell population. *J. Immunol.* **144**, 1696–1704 (1990).
40. Goulder, P. J. *et al.* Functionally inert HIV-specific cytotoxic T lymphocytes do not play a major role in chronically infected adults and children. *J. Exp. Med.* **192**, 1819–1832 (2000).
41. Kagi, D. *et al.* Cytotoxicity mediated by T cells and natural killer cells is greatly impaired in perforin-deficient mice. *Nature* **369**, 31–37 (1994).
42. Guidotti, L. G. *et al.* Intracellular inactivation of the hepatitis B virus by cytotoxic T lymphocytes. *Immunity* **4**, 25–36 (1996).
43. Guidotti, L. G. *et al.* Viral clearance without destruction of infected cells during acute HBV infection. *Science* **284**, 825–829 (1999).
44. Ramsay, A. J., Ruby, J. & Ramshaw, I. A. A case for cytokines as effector molecules in the resolution of virus infections. *Immunol. Today* **14**, 155–158 (1993).
45. Price, P. *et al.* Cytotoxic CD8<sup>+</sup> T lymphocytes reactive with human immunodeficiency virus-1 produce granulocyte/macrophage colony-stimulating factor and variable amounts of interleukins 2, 3, and 4 following stimulation with the cognate epitope. *Clin. Immunol. Immunopathol.* **74**, 100–106 (1995).
46. Jasoy, C. *et al.* Human immunodeficiency virus type 1-specific cytotoxic T lymphocytes release gamma interferon, tumor necrosis factor alpha (TNF-alpha), and TNF-beta when they encounter their target antigens. *J. Virol.* **67**, 2844–2852 (1993).
47. Meylan, P. R., Guatelli, J. C., Munis, J. R., Richman, D. D. & Kornbluth, R. S. Mechanisms for the inhibition of HIV replication by interferons- $\alpha$ ,  $\beta$ , and  $\gamma$  in primary human macrophages. *Virology* **193**, 138–148 (1993).
48. Emilie, D., Maillot, M. C., Nicolas, J. F., Fior, R. & Galanaud, P. Antagonistic effect of interferon-gamma on tat-induced transactivation of HIV long terminal repeat. *J. Biol. Chem.* **267**, 20565–20570 (1992).
49. Bollinger, R. C. *et al.* Cytokines from vaccine-induced HIV-1 specific cytotoxic T lymphocytes: effects on viral replication. *AIDS Res. Hum. Retroviruses* **9**, 1067–1077 (1993).
50. Harter, T., Jasoy, C., Harter, E., Johnson, R. P. & Walker, B. D. Induction of HIV-1 replication in a chronically infected T-cell line by cytotoxic T lymphocytes. *J. Acquir. Immune Defic. Syndr.* **6**, 865–871 (1993).
51. Wagner, L. *et al.*  $\beta$ -Chemokines are released from HIV-1-specific cytolytic T-cell granules complexed to proteoglycans. *Nature* **391**, 908–911 (1998).
52. Price, D. A. *et al.* Antigen-specific release of  $\beta$ -chemokines by anti-HIV-1 cytotoxic T lymphocytes. *Curr. Biol.* **8**, 355–358 (1998).
53. Cocchi, F. *et al.* Identification of RANTES, MIP-1 $\alpha$ , and MIP-1 $\beta$  as the major HIV-suppressive factors produced by CD8<sup>+</sup> T cells. *Science* **270**, 1811–1815 (1995).
54. Mackewicz, C. & Levy, J. A. CD8<sup>+</sup> cell anti-HIV activity: nonlytic suppression of virus replication. *AIDS Res. Hum. Retroviruses* **8**, 1039–1050 (1992).
55. Levy, J. A., Mackewicz, C. E. & Barker, E. Controlling HIV pathogenesis: the role of noncytotoxic anti-HIV response of CD8<sup>+</sup> T cells. *Immunol. Today* **17**, 217–224 (1996).
56. Copeland, K., McKay, P. J. & Rosenthal, K. L. Suppression of activation of the HIV LTR by CD8<sup>+</sup> cells is not lentivirus specific. *AIDS Res. Hum. Retroviruses* **11**, 1321–1325 (1995).
57. Yang, O. O. *et al.* Efficient lysis of human immunodeficiency virus type 1-infected cells by cytotoxic T lymphocytes. *J. Virol.* **70**, 5799–5806 (1996).
58. Collins, K. L., Chen, B. K., Kalams, S. A., Walker, B. D. & Baltimore, D. HIV-1 Nef protein protects infected primary cells against killing by cytotoxic T lymphocytes. *Nature* **391**, 397–401 (1998).
59. Klenerman, P. *et al.* Cytotoxic T lymphocytes and viral turnover in HIV type 1 infection. *Proc. Natl Acad. Sci. USA* **93**, 15323–15328 (1996).
60. Shankar, P., Xu, Z. & Lieberman, J. Viral-specific cytotoxic T lymphocytes lyse human immunodeficiency virus-infected primary T lymphocytes by the granule exocytosis pathway. *Blood* **94**, 3084–3093 (1999).
61. Hadida, F. *et al.* Cutting edge: RANTES regulates Fas ligand expression and killing by HIV-specific CD8 cytotoxic T cells. *J. Immunol.* **163**, 1105–1109 (1999).
62. Zajac, A. J. *et al.* Viral immune evasion due to persistence of activated T cells without effector function. *J. Exp. Med.* **188**, 2205–2213 (1998).
63. Kalams, S. A. & Walker, B. D. The critical need for CD4 help in maintaining effective cytotoxic T lymphocyte responses. *J. Exp. Med.* **188**, 2199–2204 (1998).
64. Appay, V. *et al.* HIV-specific CD8<sup>+</sup> T-cells produce antiviral cytokines but are impaired in cytolytic function. *J. Exp. Med.* **192**, 63–75 (2000).
65. Andersson, J. *et al.* Perforin is not co-expressed with granzyme A within cytotoxic granules in CD8 T lymphocytes present in lymphoid tissue during chronic HIV infection. *AIDS* **13**, 1295–1303 (1999).
66. Trimble, L. A. & Lieberman, J. Circulating CD8 T lymphocytes in human immunodeficiency virus-infected individuals have impaired function and downmodulate CD3 $\zeta$ , the signaling chain of the T-cell receptor complex. *Blood* **91**, 585–594 (1998).
67. Hamann, D. *et al.* Phenotypic and functional separation of memory and effector human CD8<sup>+</sup> T cells. *J. Exp. Med.* **186**, 1407–1418 (1997).
68. Champagne, P. *et al.* Skewed maturation of memory HIV-1 specific CD8 T lymphocytes. *Nature* **410**, 106–111 (2001).
69. Clerici, M. *et al.* Detection of three distinct patterns of T helper cell dysfunction in asymptomatic, human immunodeficiency virus-seropositive patients. Independence of CD4<sup>+</sup> cell numbers and clinical staging. *J. Clin. Invest.* **84**, 1892–1899 (1989).
70. Rosenberg, E. S. *et al.* Vigorous HIV-1-specific CD4<sup>+</sup> T-cell responses associated with control of viremia. *Science* **278**, 1447–1450 (1997).
71. Oxenius, A. *et al.* Early highly active antiretroviral therapy for acute HIV-1 infection preserves immune function of CD8<sup>+</sup> and CD4<sup>+</sup> T lymphocytes. *Proc. Natl Acad. Sci. USA* **97**, 3382–3387 (2000).
72. Pitcher, C. J. *et al.* HIV-1-specific CD4<sup>+</sup> T cells are detectable in most individuals with active HIV-1 infection, but decline with prolonged viral suppression. *Nature Med.* **5**, 518–525 (1999).
73. Geijtenbeek, T. B. *et al.* DC-SIGN, a dendritic cell-specific HIV-1-binding protein that enhances trans-infection of T cells. *Cell* **100**, 587–597 (2000).
74. Pope, M., Gezelter, S., Gallo, N., Hoffman, L. & Steinman, R. M. Low levels of HIV-1 infection in cutaneous dendritic cells promote extensive viral replication upon binding to memory CD4<sup>+</sup> T cells. *J. Exp. Med.* **182**, 2045–2056 (1995).
75. Rosenberg, E. S. *et al.* Immune control of HIV-1 after early treatment of acute infection. *Nature* **407**, 523–526 (2000).
76. Ridge, J. P., Di Rosa, F. & Matzinger, P. A conditioned dendritic cell can be a temporal bridge

- between a CD4<sup>+</sup> T-helper and a T-killer cell. *Nature* **393**, 474–478 (1998).
77. Walter, E. A. *et al.* Reconstitution of cellular immunity against cytomegalovirus in recipients of allogeneic bone marrow by transfer of T-cell clones from the donor. *N. Engl. J. Med.* **333**, 1038–1044 (1995).
  78. Brodie, S. J. *et al.* In vivo migration and function of transferred HIV-1-specific cytotoxic T cells. *Nature Med.* **5**, 34–41 (1999).
  79. Tan, R. *et al.* Rapid death of adoptively transferred T cells in acquired immunodeficiency syndrome. *Blood* **93**, 1506–1510 (1999).
  80. Kanazawa, S., Okamoto, T. & Peterlin, B. M. Tat competes with CIITA for the binding to P-TEFb and blocks the expression of MHC class II genes in HIV infection. *Immunity* **12**, 61–70 (2000).
  81. Wodarz *et al.* *Proc. R. Soc. Lond. B* (in the press).
  82. McMichael, A. J. *et al.* Memory CD8<sup>+</sup> T cells in HIV infection. *Phil. Trans. R. Soc. Lond. B* **355**, 363–367 (2000).
  83. Koenig, S. *et al.* Transfer of HIV-1 specific cytotoxic T lymphocytes to an AIDS patient leads to selection for mutant HIV variants and subsequent disease progression. *Nature Med.* **1**, 330–336 (1995).
  84. Borrow, P. *et al.* Antiviral pressure exerted by HIV-1-specific cytotoxic T lymphocytes (CTLs) during primary infection demonstrated by rapid selection of CTL escape virus. *Nature Med.* **3**, 205–211 (1997).
  85. Price, D. A. *et al.* Positive selection of HIV-1 cytotoxic T lymphocyte escape variants during primary infection. *Proc. Natl Acad. Sci. USA* **94**, 1890–1895 (1997).
  86. Goulder, P. J. *et al.* Late escape from an immunodominant cytotoxic T-lymphocyte response associated with progression to AIDS. *Nature Med.* **3**, 212–217 (1997).
  87. Kelleher, A. D. *et al.* Clustered mutations in HIV-1 gag are consistently required for escape from HLA-B27-restricted CTL responses. *J. Exp. Med.* **193**, 375–386 (2001).
  88. Berthet-Colominas, C. *et al.* Head-to-tail dimers and interdomain flexibility revealed by the crystal structure of HIV-1 capsid protein (p24) complexed with a monoclonal antibody Fab. *EMBO J.* **18**, 1124–1136 (1999).
  89. Jones, I. M. & Morikawa, Y. The molecular basis of HIV capsid assembly. *Rev. Med. Virol.* **8**, 87–95 (1998).
  90. Zhang, W. H., Hockley, D. J., Nermut, M. V. & Jones, I. M. Functional consequences of mutations in HIV-1 Gag p55 selected by CTL pressure. *Virology* **203**, 101–105 (1994).
  91. Goulder, P. J. R. *et al.* Patterns of immunodominance in HIV-1-specific cytotoxic T lymphocyte responses in two human histocompatibility leukocyte antigens (HLA)—identical Siblings with HLA-A\*0201 are influenced by epitope mutation. *J. Exp. Med.* **185**, 1423–1433 (1997).
  92. Allen, T. M. *et al.* Tat-specific cytotoxic T lymphocytes select for SIV escape variants during resolution of primary viraemia. *Nature* **407**, 386–390 (2000).
  93. Nowak, M. A. *et al.* Antigenic oscillations and shifting immunodominance in HIV-1 infections. *Nature* **375**, 606–611 (1995).
  94. Janssens, W., Buve, A. & Nkengasong, J. N. The puzzle of HIV-1 subtypes in Africa. *AIDS* **11**, 705–712 (1997).
  95. Le Gall, S. *et al.* Nef interacts with the mu subunit of clathrin adaptor complexes and reveals a cryptic sorting signal in MHC I molecules. *Immunity* **8**, 483–495 (1998).
  96. Cohen, G. B. *et al.* The selective downregulation of class I major histocompatibility complex proteins by HIV-1 protects HIV-infected cells from NK cells. *Immunity* **10**, 661–671 (1999).
  97. Albert, M. L., Sauter, B. & Bhardwaj, N. Dendritic cells acquire antigen from apoptotic cells and induce class I-restricted CTLs. *Nature* **392**, 86–89 (1998).
  98. Xu, X.-N. *et al.* Evasion of CTL responses by nef-dependent induction of Fas ligand expression on SIV-infected cells. *J. Exp. Med.* **186**, 7–16 (1997).
  99. Xu, X. N. *et al.* Induction of Fas ligand expression by HIV involves the interaction of Nef with the T cell receptor  $\zeta$  chain. *J. Exp. Med.* **189**, 1489–1496 (1999).
  100. McMichael, A. J. & Bangham, C. R. M. (eds) *Semin. Virol.* **7**, 1 (1996).
  101. Connor, R. I. *et al.* Immunological and virological analyses of persons infected by human immunodeficiency virus type 1 while participating in trials of recombinant gp120 subunit vaccines. *J. Virol.* **72**, 1552–1576 (1998).
  102. Barouch, D. H. *et al.* Augmentation of immune responses to HIV-1 and simian immunodeficiency virus DNA vaccines by IL-2/Ig plasmid administration in rhesus monkeys. *Proc. Natl Acad. Sci. USA* **97**, 4192–4197 (2000).
  103. Tanchot, C. *et al.* Modifications of CD8<sup>+</sup> T cell function during in vivo memory or tolerance induction. *Immunity* **8**, 581–590 (1998).
  104. Kent, S. J. *et al.* Enhanced T-cell immunogenicity and protective efficacy of a human immunodeficiency virus type 1 vaccine regimen consisting of consecutive priming with DNA and boosting with recombinant fowlpox virus. *J. Virol.* **72**, 10180–10188 (1998).
  105. Gallimore, A. *et al.* Early suppression of SIV replication by CD8<sup>+</sup> nef-specific cytotoxic T cells in vaccinated macaques. *Nature Med.* **1**, 1167–1173 (1995).
  106. Rowland-Jones, S. L. *et al.* Cytotoxic T cell responses to multiple conserved HIV epitopes in HIV-resistant prostitutes in Nairobi. *J. Clin. Invest.* **102**, 1758–1765 (1998).
  107. Lifson, J. D. *et al.* Containment of simian immunodeficiency virus infection: cellular immune responses and protection from rechallenge following transient postinoculation antiretroviral treatment. *J. Virol.* **74**, 2584–2593 (2000).
  108. Putkonen, P., Makitalo, B., Bottiger, D., Biberfeld, G. & Thorstensson, R. Protection of human immunodeficiency virus type 2-exposed seronegative macaques from mucosal simian immunodeficiency virus transmission. *J. Virol.* **71**, 4981–4984 (1997).
  109. Allen, T. M. *et al.* Induction of AIDS virus-specific CTL activity in fresh, unstimulated peripheral blood lymphocytes from rhesus macaques vaccinated with a DNA prime/modified vaccinia virus Ankara boost regimen. *J. Immunol.* **164**, 4968–4978 (2000).
  110. Mortara, L. *et al.* Selection of virus variants and emergence of virus escape mutants after immunization with an epitope vaccine. *J. Virol.* **72**, 1403–1410 (1998).

# Pathways to neuronal injury and apoptosis in HIV-associated dementia

Marcus Kaul\*, Gwenn A. Garden† & Stuart A. Lipton\*

\*Center for Neuroscience and Aging Research, The Burnham Institute, 10901 North Torrey Pines Road, La Jolla, California 92037, USA (e-mail: mkaul@burnham.org and slipton@burnham.org)

†Department of Neurology, University of Washington, Seattle, Washington 98195, USA (email: gagarden@u.washington.edu)

**Human immunodeficiency virus-1 (HIV-1) can induce dementia with alarming occurrence worldwide. The mechanism remains poorly understood, but discovery in brain of HIV-1-binding sites (chemokine receptors) provides new insights. HIV-1 infects macrophages and microglia, but not neurons, although neurons are injured and die by apoptosis. The predominant pathway to neuronal injury is indirect through release of macrophage, microglial and astrocyte toxins, although direct injury by viral proteins might also contribute. These toxins overstimulate neurons, resulting in the formation of free radicals and excitotoxicity, similar to other neurodegenerative diseases. Recent advances in understanding the signalling pathways mediating these events offer hope for therapeutic intervention.**

**T**he syndrome of cognitive and motor dysfunction observed after infection with human HIV-1 has been designated HIV-associated dementia (HAD). Although highly active antiretroviral therapy (HAART) has resulted in a decrease in the incidence of HAD<sup>1,2</sup>, it does not seem to provide complete protection from or reversal of HAD<sup>3,4</sup>, and the prevalence of the dementia may eventually increase as people live longer with AIDS<sup>5,6</sup>. Currently there is no specific treatment for HAD, mainly because of an incomplete understanding of how HIV infection causes neuronal injury and apoptosis. The principal pathway for HIV entry into the central nervous system (CNS) is through infected monocytes. The predominant pathogenesis of HAD is believed to involve activation of monocytic cells (macrophages and microglia) and their subsequent release of toxins that lead to neuronal and astrocytic dysfunction. Macrophages and microglia can be activated by HIV infection itself, by interaction with viral proteins, or by immune stimulation due to concurrent infection or other factors<sup>7</sup>. It is possible that direct effects of viral proteins on neurons may also contribute to neurodegeneration, although neurons themselves are not infected by HIV-1. Here we review the extracellular and intracellular signalling pathways leading to macrophage or microglial activation, as well as those induced in neurons and astrocytes. These pathways are of potential therapeutic importance as targets for the prevention or treatment of HAD.

The prevalence of HAD was estimated in the early 1990s to be as high as 20–30% of those patients with advanced HIV disease and low CD4 counts<sup>8</sup>. Many experts believe that HAD is now the most common cause of dementia worldwide among people aged 40 or less, and it is a significant independent risk factor for death due to AIDS<sup>9</sup>. With the advent of HAART, the incidence of HAD decreased to as low as 10.5% (ref. 10), but in recent years its incidence as an AIDS-defining illness has actually increased<sup>3</sup>. The proportion of new cases of HAD demonstrating a CD4 count greater than 200  $\mu\text{l}^{-1}$  is also increasing<sup>10</sup>. Moreover, a less fulminant form of neurological dysfunction, termed minor cognitive/motor disorder (MCMD), may be more prevalent than frank dementia in the HAART era, and remains a significant

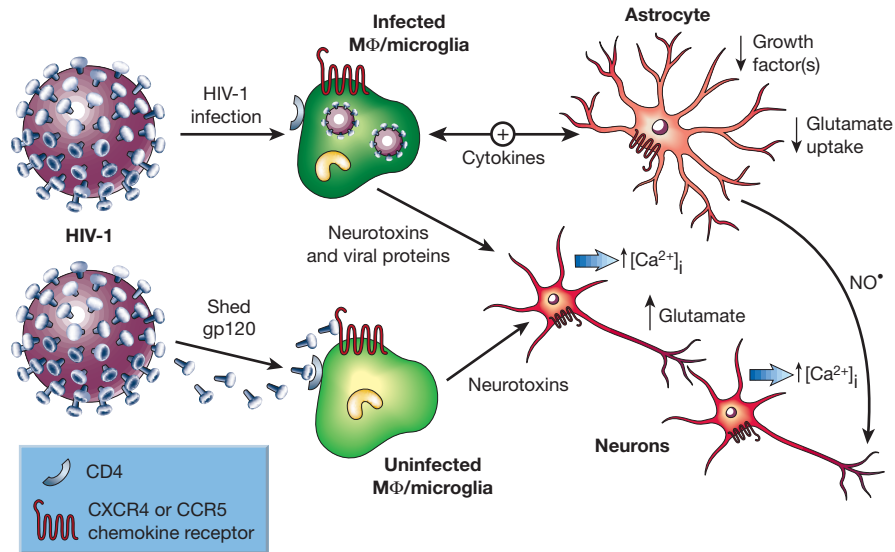
independent risk factor for AIDS mortality<sup>9</sup>. These findings support the hypothesis that HAART does not provide complete protection from the development of HAD. Unfortunately, HIV protease inhibitors and several of the nucleoside analogues penetrate poorly into the CNS<sup>11</sup>, allowing early CNS infection to evolve independently over time in a protected brain reservoir. Thus, despite the finding that improved systemic control of viral replication is associated with a decrease in the incidence of HAD, whether this effect will be long lasting is open to question<sup>4,12</sup>. In fact, distinct antiviral drug-resistance patterns in the plasma and cerebrospinal fluid (CSF) compartments have recently been reported<sup>13</sup>. It is possible that the proportion of HIV-infected individuals who develop disability secondary to HAD will increase as improvements in control of peripheral viral replication and the treatment of opportunistic infections continue to extend life expectancy, resulting in an increase in the overall prevalence of HAD. A more complete understanding of the pathogenesis of HAD will help identify therapeutic targets for its prevention and treatment.

## HIV entry into the brain and initiation of HAD

HIV enters the CNS early in the course of infection, and the virus resides primarily in microglia and macrophages. However, infection of these cells may not be sufficient to initiate neurodegeneration<sup>12</sup>. It has been proposed that factors associated with advanced HIV infection in the periphery (non-CNS) are important triggers for events leading to dementia. One such factor could be the increased number of circulating monocytes that express CD16 and CD69. These activated cells adhere to the normal endothelium of the brain microvasculature, transmigrate, and then trigger a number of deleterious processes<sup>12</sup>.

The blood–brain barrier (BBB) is crucial in HIV infection of the CNS<sup>12,14</sup>. Microglial and astrocytic chemokines (cell migration (chemotaxis)-inducing cytokines), such as monocyte chemoattractant protein (MCP)-1, seem to regulate migration of peripheral blood mononuclear cells through the BBB<sup>14</sup>. Histological studies in specimens from HIV-1-infected humans and simian immunodeficiency virus (SIV)-infected rhesus macaques show that lymphocytes and monocytes migrate into the brain<sup>15,16</sup>. However, the pathophysiological relevance of CNS-invading





**Figure 1** Current model of HIV-related neuronal damage involving cell–cell signalling. One of the main pathways of HIV-1 entry into the brain occurs by means of infected macrophages (MΦ). Once in the brain, infected macrophages or microglia release viral envelope proteins (gp120), cytokines (for example, TNF-α) and chemokines, which in turn activate uninfected macrophages and microglia. Immune activated- and HIV-infected brain macrophages/microglia release potentially neurotoxic substances. These substances include quinolinic acid and other EAAs such as glutamate and L-cysteine, arachidonic acid, PAF, NTox, free radicals and TNF-α. These substances induce neuronal injury, dendritic and synaptic damage, and apoptosis. Macrophages express surface chemokine receptors (CCR5 and CXCR4) and CD4, which serve as co-receptors for HIV/gp120. Many neurons and astrocytes also express CXCR4 and CCR5, raising the possibility of direct interaction with gp120. Macrophages and astrocytes have mutual feedback loops (signified by the reciprocal arrows). Immune stimulation by inflammatory cytokines, such as TNF-α and IL-1β, participate in this intercellular network in several ways. For example, HIV infection or gp120 stimulation of macrophages enhances their production of TNF-α and IL-1β. These cytokines stimulate astrocytosis. Reactive astrocytes may then release the free-radical nitric oxide (NO\*), which in turn may react with superoxide to form the neurotoxic molecule peroxynitrite. Arachidonate released from macrophages impairs astrocyte clearing of the neurotransmitter glutamate and thus contributes to excitotoxicity. Neuronal injury is mediated predominantly by overactivation of NMDAR-coupled ion channels that allow excessive influx of Ca<sup>2+</sup>. This in turn triggers a variety of potentially harmful enzymes, free-radical formation and release of glutamate. Glutamate subsequently overstimulates NMDARs on neighbouring neurons, initiating further injury. This final common pathway to neurotoxicity can be blocked by NMDAR antagonists. For certain neurons, depending on their exact repertoire of ionic channels, this form of damage can also be ameliorated by voltage-activated calcium channel antagonists or non-NMDA glutamate antagonists. In addition, β-chemokine-receptor agonists can confer partial protection against neuronal damage induced by HIV/gp120 or excessive NMDAR activation.

lymphocytes in HAD is not clear<sup>16</sup>. Cellular migration also involves adhesion molecules, and increased expression of vascular cell-adhesion molecule-1 (VCAM-1) has been implicated in mononuclear cell migration into the brain during HIV and SIV infection<sup>17</sup>. It has also been suggested that the inflammatory cytokine tumour-necrosis factor-α (TNF-α) opens a paracellular route for HIV-1 across the BBB<sup>18</sup>. These findings indicate that one reason why HAD rarely occurs before the onset of advanced HIV disease is that a vicious cycle of immune dysregulation and BBB dysfunction is required to achieve sufficient entry of infected or activated immune cells into the brain to cause neuronal injury. However, alterations of the BBB have been observed in transgenic mice expressing the HIV envelope protein gp120 in a form that circulates in plasma<sup>19</sup>, suggesting that circulating virus or envelope proteins may also cause BBB dysfunction during the viremic phase of primary infection.

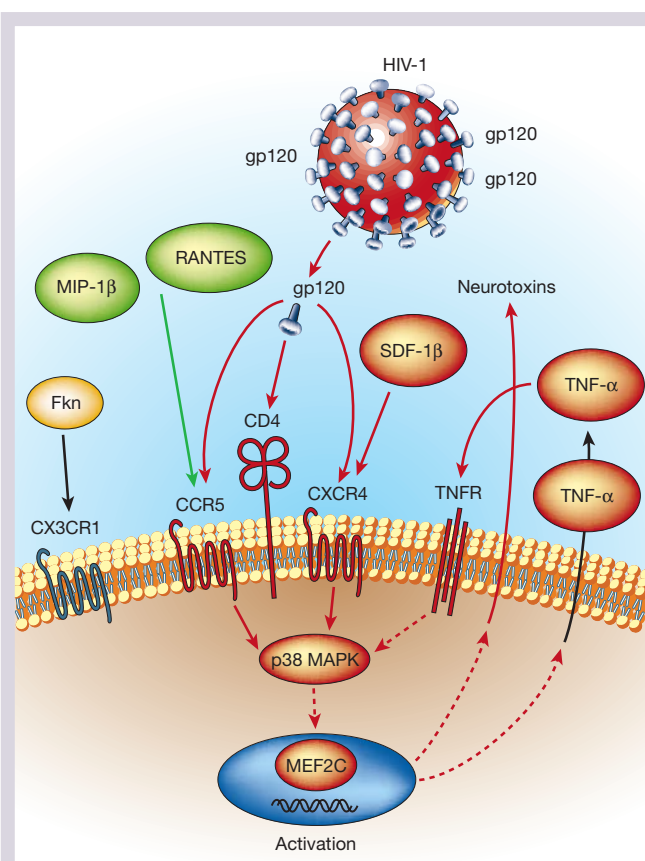
HIV infection within the CNS can be detected by measurements of viral RNA in CSF. Several but not all studies have reported a positive correlation between CSF viral load and the degree of cognitive dysfunction in patients with HAD or MCMD<sup>20–22</sup>. CSF viral load also seems to correlate with viral load in brain measured by quantitative polymerase chain reaction<sup>22,23</sup>, and the highest concentrations of virus are observed in those subcortical structures most frequently affected in patients with HAD<sup>23</sup>.

### Pathology and pathogenesis of HAD

The neuropathology associated with HIV infection in the brain, termed HIV encephalitis, is characterized by widespread reactive astrocytosis, myelin pallor, and infiltration predominantly by monocyteoid cells, including blood-derived macrophages, resident

microglia and multinucleated giant cells. However, numbers of HIV-infected cells, multinucleated giant cells or viral antigen in CNS tissue do not correlate well with measures of cognitive function<sup>24,25</sup>. The pathological features most closely associated with the clinical signs of HAD include increased numbers of microglia<sup>24</sup>, elevated TNF-α messenger RNA in microglia and astrocytes<sup>26</sup>, evidence of excitotoxins<sup>27,28</sup>, decreased synaptic and dendritic density<sup>25,29</sup>, and selective neuronal loss<sup>30,31</sup>. Several groups have demonstrated that HAD is associated with evidence of neuronal apoptosis<sup>32–34</sup>, but this finding is not clearly associated with viral burden<sup>32</sup> or a history of dementia<sup>35</sup>. The topographic distribution of neuronal apoptosis is correlated with evidence of structural atrophy and closely associated with markers of microglial activation, especially within subcortical deep grey structures<sup>35</sup>, which may show a predilection for atrophy in HAD.

The neuropathology observed in HAD, coupled with extensive research on both *in vitro* and animal models of HIV-induced neurodegeneration, have led to a complicated model for the pathogenesis of HAD. As with other emerging models of neurodegenerative disease, it is likely that a construct similar to the multi-hit model of oncogenesis will be the most effective way to understand all of the factors involved in the pathogenesis of HAD. Figure 1 shows the potential intercellular interactions that can lead to neuronal injury in the setting of HIV infection. Macrophages and microglia can be infected by HIV-1, but they can also be activated by factors released from infected cells, including cytokines and shed viral proteins such as gp120. Microglial activation affects all cell types in the CNS, resulting in upregulation of cytokines, chemokines and endothelial adhesion molecules<sup>7,12</sup>. Some of these molecules may contribute to neuronal damage and apoptosis through direct or indirect routes. In addition, activated microglia



**Figure 2** Microglial/macrophage signalling in HAD. HIV/gp120 (gp) interacts with chemokine receptors CXCR4 or CCR5 in conjunction with CD4 to stimulate or infect (if the entire virus is present) microglia and macrophages. Natural ligands of CXCR4 and CCR5, the  $\alpha$ -chemokine SDF-1 and the  $\beta$ -chemokines MIP-1 $\beta$  and RANTES, respectively, interfere with HIV/gp120 binding and signalling. However, only the  $\beta$ -chemokines can prevent the neurotoxic effect of activated microglia and macrophages. The  $\delta$ -chemokine fractalkine (Fkn) mediates communication between neurons and microglia, which are activated by Fkn. The viral envelope protein gp120 triggers a signalling pathway that involves p38 MAPK, a pivotal factor in immune stimulation of macrophages that activates the transcription factor MEF2C. HIV/gp120 induces the release of neurotoxic substances, including EAs, arachidonic acid and related molecules such as PAF, which engenders neuronal glutamate release; gp120 also induces release of inflammatory cytokines, such as TNF- $\alpha$ . Inflammatory cytokines can activate adjacent microglia/macrophages and astrocytes, and thus contribute indirectly to brain injury.

release excitatory amino acids (EAAs) and related substances, including glutamate, quinolate, cysteine and the amine NTx<sup>36–40</sup>. EAAs released by infected or activated microglia can induce neuronal apoptosis through a process known as excitotoxicity, which engenders excessive  $\text{Ca}^{2+}$  influx and free radical (nitric oxide and superoxide anion) formation by overstimulation of glutamate receptors<sup>41</sup>. Some HIV proteins, such as gp120 and Tat, have also been reported to be directly neurotoxic, although high concentrations of viral protein may be needed or neurons may have to be cultured in isolation to see these direct effects<sup>42,43</sup>. Importantly, toxic viral proteins and factors released from microglia may act synergistically to promote neurodegeneration, even in the absence of extensive viral invasion of the CNS.

### The role of chemokine receptors in HAD

Entry of HIV-1 into cells such as T cells or macrophages/microglia occurs after binding of the viral envelope protein gp120 to chemokine receptors in conjunction with CD4. Generally, T cells are infected through the  $\alpha$ -chemokine receptor CXCR4 and/or the  $\beta$ -chemokine receptor CCR5. In contrast, monocytes, macrophages

and microglia are infected primarily through CCR5 or CCR3 (but CXCR4 may also be involved)<sup>44,45</sup>. Chemokine receptors can also be present on neurons and astrocytes in the brain<sup>46</sup>, although these cells are not thought to harbour productive infection. *In vitro* studies indicate that distinct chemokine receptors mediate HIV-associated neuronal damage, whereas others may serve a protective role<sup>42,47</sup>. Three chemokine receptors (with their respective ligands in parentheses) are of particular interest in HAD: CXCR4 (stromal-derived factor (SDF)-1 $\alpha/\beta/\gamma$ ); CCR5 (RANTES or MIP-1 $\alpha/\beta$ ); and CX3CR1 (fractalkine). CXCR4 is expressed on neurons, microglia, astrocytes and endothelia in the brain<sup>48–50</sup>. SDF-1 is produced by astrocytes, macrophages, neurons and Schwann cells, and an increase in SDF-1 mRNA has been detected in HIV encephalitis<sup>51</sup>.

Chemokine receptors are seven membrane-spanning, G-protein-coupled receptors, and as such trigger intracellular signalling events. For example, SDF-1 $\alpha$  can modulate synaptic transmission in the rat cerebellum by increasing the intracellular  $\text{Ca}^{2+}$  concentration<sup>52</sup>. In isolated hippocampal neurons, SDF-1 $\alpha$  causes  $\text{Ca}^{2+}$ - and cyclic AMP-dependent activation of the cAMP response element-binding protein (CREB), a key transcription factor regulating gene expression<sup>42</sup>. SDF-1 $\alpha$  also increases the activity of extracellular signal-regulated kinase-1 and -2 (ERK-1, -2) in neurons, similar to gp120 from HIV-1<sub>IIIb</sub>, a CXCR4 (or X4)-preferring strain of the virus<sup>53</sup>. SDF-1 $\alpha/\beta$ , HIV-1 virions and supernatants of HIV-infected monocyte-derived macrophages reportedly signal by a  $G_i$ -dependent decrease in cAMP and an increase in inositol-1,4,5-trisphosphate and  $[\text{Ca}^{2+}]_i$ ; these signals correlate with enhanced synaptic transmission, activation of caspase-3 and neuronal apoptosis<sup>48</sup>.

In cerebrocortical neurons and neuronal cell lines, picomolar concentrations of X4-preferring or dualtropic gp120 (as well as intact virus) can induce neurotoxicity through CXCR4 receptors<sup>47,54,55</sup>. In a study of mixed neuronal/glial cultures that mimic the cellular composition of the intact brain<sup>47</sup>, this apoptotic death seems to be mediated predominantly by the release of microglial toxins rather than by direct neuronal damage. However, nanomolar concentrations of SDF-1 $\alpha/\beta$  interacting with CXCR4 can induce apoptotic death of cerebrocortical neurons in the absence of microglial signalling, implying a direct interaction with neurons<sup>47</sup>. Interestingly, inhibition of p38 mitogen-activated protein kinase (MAPK) prevents the neurotoxicity of both gp120 and SDF-1 in these mixed cultures (Fig. 2)<sup>47</sup>. In contrast to these findings, somewhat higher concentrations of SDF-1 $\alpha$  have been reported to provide neuroprotection from X4-preferring gp120-induced damage of isolated hippocampal neurons<sup>42</sup>. Clearly, the results obtained on isolated neurons may be different from those observed in mixed neuronal/glial cultures that more closely resemble the repertoire of cells found in the brain.

The  $\beta$ -chemokine receptor CCR5 is expressed by neurons, microglia and astrocytes<sup>46</sup>, and its role in HAD seems to be distinct from the effect of the  $\alpha$ -chemokine receptor CXCR4. Although individuals lacking functional CCR5 are at reduced risk for HIV infection, *in vitro* studies have shown that activation of this chemokine receptor by RANTES or MIP-1 $\alpha/\beta$  protects neurons from gp120-induced apoptosis<sup>42,47</sup>. In addition, HIV-infected patients with higher CSF concentrations of  $\beta$ -chemokines MIP-1 $\alpha/\beta$  and RANTES perform better on neuropsychological measures than those with low or undetectable levels of  $\beta$ -chemokines<sup>56</sup>. These findings indicate that once HIV infection is established, distinct  $\beta$ -chemokines may actually be protective against progression to HAD. But other  $\beta$ -chemokines, such as MCP-1, are elevated in the CSF of patients with HAD and may correlate with CSF viral load<sup>57,58</sup>.

CX3CR1 is a chemokine receptor found on microglia and possibly neurons, and has been reported to interact with HIV-1 *in vitro*<sup>59</sup>. Fractalkine, a chemokine expressed in both a membrane-bound and soluble form by neurons, binds to CX3CR1, possibly affording neuronal–microglial interaction<sup>14</sup>. Fractalkine and CX3CR1 have been found to be upregulated in paediatric patients with HIV encephalitis, and reportedly mediate both macrophage recruitment

and neuroprotection from gp120 or Tat, at least *in vitro*<sup>60,61</sup>. These two actions would seemingly exert opposing effects on HAD. In rat microglia or hippocampal neurons, fractalkine signalling increases intracellular  $\text{Ca}^{2+}$ , activates the survival signalling molecule Akt, and triggers microglial chemotaxis<sup>61,62</sup>. A potential role of fractalkine in HAD is indicated by its cleavage from the neuronal cell surface in response to excitotoxic injury<sup>63</sup>.

### Macrophages/microglia in HAD

Macrophages and microglia are crucial in HAD because they are the only resident cells that can be productively infected with HIV-1 in the CNS<sup>7</sup>, although a non-productive or latent infection of astrocytes has been observed<sup>64</sup>. HIV-1-infected macrophages (and possibly lymphocytes) migrate into the brain and constitute the principal route of viral entry into the CNS (reviewed in ref. 12). HIV-infected or immune-stimulated macrophages/microglia produce neurotoxins, and macrophages/microglia are required for HIV-1- or gp120-induced neurotoxicity<sup>36,37,65</sup>. Macrophage/microglia damage neurons by releasing excitotoxic substances that produce excessive activation of glutamate receptors, primarily of the *N*-methyl-D-aspartate subtype (NMDAR). In addition, indirect neurotoxicity is probably mediated by macrophage- and microglial-derived inflammatory cytokines, such as interleukin (IL)-1 $\beta$  and TNF- $\alpha$ , arachidonate and its metabolites including platelet-activating factor (PAF), free radicals chemokines and viral proteins (reviewed in ref. 7). Chemokine and cytokine signalling in microglia promote p38 MAPK activity that in turn phosphorylates/activates the transcription factor MEF2C (Fig. 2). Pharmacological inhibition of p38 MAPK prevents microglial induction of TNF- $\alpha$  and inducible nitric oxide synthase (iNOS) gene expression in response to inflammatory stimuli<sup>66</sup>. The p38 signal-transduction pathway seems to be involved in mediating both microglial activation and neuronal injury<sup>7,47</sup>, indicating that modulation of this kinase might be an important therapeutic target for the prevention of HAD.

### The role of astrocytes in HAD

Astrocytosis (proliferation of activated astrocytes) as well as occasional astrocyte apoptosis occur in association with HAD, and are also observed in the gp120 transgenic mouse<sup>34,67,68</sup>. We have found an increase in the number of astrocytes labelled with antibodies against activated caspase-3 in post-mortem cerebrocortical tissue from human patients with HAD, suggesting ongoing injury to astrocytes in the setting of HIV (G.A.G. and S.A.L., unpublished data).

Astrocytes may contribute to the production or maintenance of excitotoxins like glutamate in several ways (Fig. 3). For example, the normal re-uptake of glutamate by astrocytes is impaired and release of astrocytic glutamate is induced by several factors derived from activated macrophages/microglia, including arachidonic acid and TNF- $\alpha$ <sup>69,70</sup>. Stimulation of metabotropic (G-protein-linked) glutamate receptors (mGluRs) on astrocytes may lead to increased  $[\text{Ca}^{2+}]$ , and further release of glutamate<sup>71</sup>. In addition to contributing to dysregulation of EAA homeostasis, astrocytes may be important in relaying or amplifying neurotoxic signals that emanate from activated or HIV-infected microglia. First, the HIV protein Tat can induce astrocytic expression of the  $\beta$ -chemokine MCP-1 (ref. 72). Secreted MCP-1 is a chemoattractant factor for monocytes/microglia and may recruit these cells into an environment where they subsequently release cytokines and EAAs. Second, cytokines and viral proteins promote the induction of iNOS within astrocytes<sup>73</sup>. The nitric oxide (NO) thus released may then react with superoxide anion ( $\text{O}_2^-$ ) to form neurotoxic peroxynitrite ( $\text{ONOO}^-$ ), similar to the reaction that can occur within neurons after excessive NMDAR stimulation.

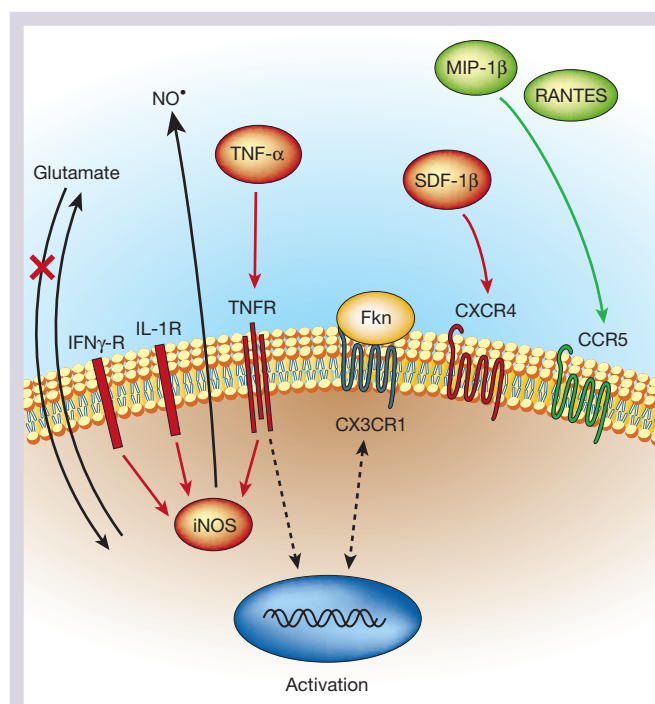
### Neuronal injury and apoptosis in HAD

Although there is general agreement that HIV does not infect neurons, the primary cause of neuronal injury remains in question. Evidence supports multiple theories for neuronal injury by various viral proteins, including Tat, Nef, Vpr and the Env proteins gp120

and gp41. Two theories predominate and are best described as the 'direct injury' hypothesis and the 'indirect' or 'bystander effect' hypothesis. They are in no way mutually exclusive, and currently available data support a role for both theories, although an indirect form of neurotoxicity seems to have more support.

The theory that HIV proteins can injure neurons directly without requiring the intermediary function of non-neuronal cells (microglia and/or astrocytes) is supported by experiments showing that viral envelope proteins are toxic in serum-free primary neuronal culture<sup>42</sup> or in neuroblastoma cell lines<sup>74</sup>. In these experimental paradigms, the impact of neurotoxic cytokines and EAAs secreted from non-neuronal cells is minimized because serum-free neuronal cultures contain few if any non-neuronal cells, and neuroblastoma lines do not contain cells of other phenotypes. The HIV coat protein gp120 interacts with several members of the chemokine receptor family, and the direct form of HIV-induced neuronal injury may be mediated by chemokine receptor signalling. Indeed, experiments aimed at blocking chemokine receptor signalling can in some cases prevent HIV/gp120-induced neuronal apoptosis<sup>48,61</sup>. Moreover, some chemokines, such as SDF-1, seem to be directly neurotoxic through stimulation of  $\alpha$ -chemokine receptors<sup>47,55</sup>. In addition, nanomolar concentrations of gp120 have been reported to interact with the glycine-binding site of the NMDAR<sup>75</sup>, suggesting another mechanism by which HIV/gp120 may have a direct effect on neuronal cell death. However, these concentrations are high compared with the picomolar amounts of gp120 that can mediate indirect neuronal injury through microglial stimulation.

HIV/Tat can be taken up into PC-12 cells by a receptor-mediated mechanism<sup>43</sup> and may also have a direct effect on neurons by



**Figure 3** Astrocyte signalling in HAD. Astrocytes express the HIV co-receptors CXCR4 and CCR5 in addition to other chemokine receptors, but lack CD4. Therefore, astrocytic reactivity may be influenced by natural ligands of chemokine receptors. Through these chemokine receptors, astrocytes may possibly be stimulated by a CD4-independent effect of HIV/gp120. Astrocytes are activated by inflammatory cytokines, including TNF- $\alpha$ , IL-1 $\beta$  and interferon- $\gamma$  (IFN- $\gamma$ ). Exposure to arachidonic acid released from macrophages and cytokine activation results in impaired glutamate uptake, increased glutamate release, and induction of iNOS, leading to release of potentially neurotoxic nitric oxide. TNF- $\alpha$  also promotes expression of the astrocytic fractalkine (Fkn) receptor, CXCR1. Stimulation of CXCR1 on astrocytes induces release of a soluble factor that triggers microglial proliferation.



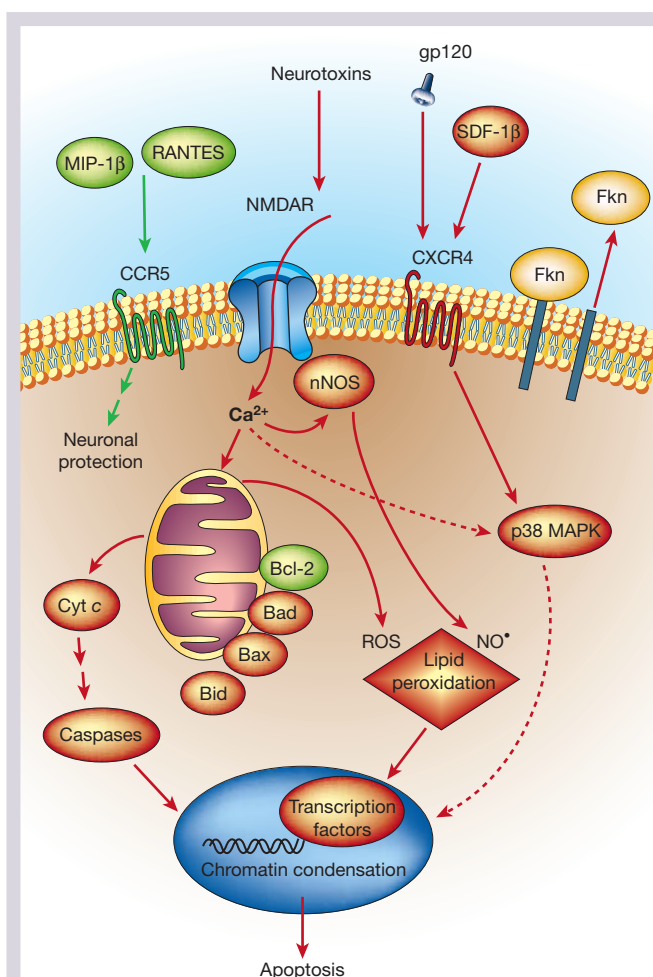
potentiating the response to excitotoxic stimuli<sup>76</sup>. Experiments using cultured hippocampal neurons revealed that HIV/Vpr may be directly neurotoxic through formation of a cation-permeable channel<sup>77</sup>. However, the meaning of all of these *in vitro* findings must be interpreted in the context of the limitations of the experimental paradigm and concentration of HIV proteins used. Most of the experimental results described above were obtained in the absence of non-neuronal cells and therefore a predominantly indirect effect would not be detected. In addition, the concentrations of HIV proteins used were frequently well above the picomolar or lower range thought to be present in brain or CSF from patients with HAD.

Apoptotic neurons do not co-localize with infected microglia in HAD patients<sup>78</sup>, supporting the hypothesis the HIV infection causes neurodegeneration through the release of soluble factors. Thus, the propensity for cell–cell interactions mandates that disease pathogenesis *in vitro* be approached in a ‘mixed’ neuronal/glial primary culture system that recapitulates the type and proportion of cells normally found in the intact brain (Fig. 1). Systems designed to study the effect of soluble factors released from microglia have included mixed cerebrocortical cultures from human fetal brain directly infected with HIV<sup>78</sup>, severe combined immunodeficiency mice inoculated with HIV-infected human monocytes<sup>79</sup>, and mixed rodent cerebrocortical cultures exposed to very low concentrations of the envelope protein HIV/gp120.

Using such model culture systems, the available data support a predominantly indirect neurotoxic effect that occurs because of the response of non-neuronal cells to HIV infection or shed HIV proteins. Much of the data stem from experiments designed to examine the toxicity of HIV envelope proteins or supernatants of infected macrophages<sup>36,65,80</sup>. Picomolar concentrations of HIV/gp120 induce injury and apoptosis in primary rodent and human neurons<sup>65,81</sup>. We found that the predominant mode of HIV/gp210 neurotoxicity to cerebrocortical neurons requires the presence of macrophages/microglia<sup>7,47</sup>. HIV-1-infected or gp120-stimulated mononuclear phagocytes release neurotoxins that stimulate the NMDAR, as described earlier. NMDAR antagonists can ameliorate neuronal cell death *in vitro* resulting from HIV-infected macrophages or purified recombinant gp120 (refs 82, 83). Transgenic mice expressing gp120 in the brain manifest neuropathological features similar to those observed in HAD cases, and these changes are also prevented by treatment with an NMDAR antagonist<sup>67,84</sup>. It should be noted, however, that rodent models are not perfect in that they are not productively infected by HIV, and for this reason SIV-infected macaques and human fetal cerebrocortical cultures have proven invaluable in validating these concepts.

Excessive stimulation of the NMDAR induces several detrimental intracellular signals that contribute to neuronal cell death by apoptosis or necrosis, depending on the intensity of the initial insult (Fig. 4)<sup>41</sup>. If the initial excitotoxic insult is severe, the cells die early from loss of ionic homeostasis, resulting in acute swelling and lysis (necrosis). If the insult is more mild, neurons enter a delayed death pathway known as apoptosis<sup>41</sup>. Neuronal apoptosis after excitotoxic insult involves  $\text{Ca}^{2+}$  overload, p38 MAPK activation, release of cytochrome *c* from mitochondria, activation of caspases, free-radical formation, lipid peroxidation and chromatin condensation<sup>85–87</sup>. Similarly, we have found that neurons exposed to HIV/gp120 and grown in mixed cerebrocortical cultures containing astrocytes and microglia demonstrate release of mitochondrial cytochrome *c*, caspase activation, chromatin condensation and apoptosis which is blocked by inhibition of the p38 MAP kinase<sup>47,88</sup>. The NMDAR antagonist memantine prevents neuronal damage in response to HIV/gp120 in cultured neurons as well as in the gp120 transgenic mouse<sup>82,84</sup>.

In addition to chemokines and EAAs, HIV-infected or gp120-activated microglia also release inflammatory cytokines, including TNF- $\alpha$  and IL-1 $\beta$ . Among other actions, both of these cytokines stimulate release of L-cysteine from macrophages, and pharmacological blockade of IL-1 $\beta$  or antibody neutralization of TNF- $\alpha$  prevents



**Figure 4** Neuronal signalling in HAD. Pathways leading to apoptosis involve excessive activation of NMDARs and  $\alpha$ -chemokine receptors. Overstimulation of the NMDAR is triggered by neurotoxins released from HIV-infected or immune-stimulated macrophages/microglia and by impaired clearance of glutamate that under normal conditions would have been taken up by astrocytes. Consequently, excessive  $\text{Ca}^{2+}$  influx into neurons triggers activation of p38 MAPK, mitochondrial  $\text{Ca}^{2+}$  overload and release of cytochrome *c* (Cyt *c*), free-radical generation (nitric oxide ( $\text{NO}^*$ ) and reactive oxygen species (ROS)), caspase activation, and apoptosis. NMDARs are physically tethered to neuronal nitric oxide synthase (nNOS), facilitating its activation. The Bcl-2 family members Bad, Bax and Bid promote apoptosis mediated by glutamate, ROS and TNF- $\alpha$ , respectively. Bcl-2 prevents apoptosis, apparently by attenuating cytochrome *c* release and ROS production. Activation of the p38 MAPK pathway by a  $\text{Ca}^{2+}$ -mediated mechanism and possibly by oxidative stress may lead to phosphorylation/activation of transcription factors involved in apoptosis. Stimulation of the  $\alpha$ -chemokine receptor CXCR4 can also induce several pathways in neurons, including activation of p38 MAPK, which leads to apoptosis. In contrast, activation of the  $\beta$ -chemokine receptor CCR5 initiates an as yet uncharacterized neuroprotective pathway that interferes with toxicity triggered by HIV/gp120 or NMDARs. The chemokine fractalkine (Fkn) is released from neurons subsequent to excitotoxic injury, and may represent feedback signalling onto non-neuronal cells.

this release<sup>40</sup>. Under physiological or pathophysiological conditions, L-cysteine can stimulate NMDARs and lead to neuronal apoptosis<sup>40</sup>. TNF- $\alpha$  is capable of stimulating apoptosis in human neurons<sup>89</sup>, but an indirect route of injury cannot be excluded. Expression of TNF- $\alpha$  and its receptor are elevated in brain from patients with HAD<sup>26</sup>. Experiments aimed at addressing the question of interactions between neurotoxins associated with HAD have revealed that TNF- $\alpha$  and HIV/Tat synergize to promote neuronal death, and this effect is prevented by antioxidants<sup>90</sup>. It remains possible that TNF- $\alpha$  can activate caspases within neurons via TNF- $\alpha$  receptor-1 (TNFR1), as

TNFR1 is found on at least some neurons, and it can trigger caspase-8 activation. Indeed, we have found that antibody neutralization of TNF- $\alpha$  or inhibition of caspase-8 prevents the neurotoxicity of HIV/gp120 in cultured cerebrocortical neurons<sup>88</sup>, and caspase-8 activity can directly or indirectly activate caspase-3, leading to apoptosis. These findings suggest that inflammatory cytokines, including TNF- $\alpha$  and IL-1 $\beta$ , may have important synergistic roles in HIV-associated neuropathology.

### Strategies for treatment or prevention of HAD

Based on the evolving pathogenesis of HAD, several potential therapeutic strategies to attenuate neuronal damage are worth exploring. The present arsenal of antiretroviral drugs are not effective in penetrating the BBB, so adjunctive measures are being developed and tested. Among others, agents warranting consideration include NMDAR blockers, chemokines, chemokine- and cytokine-receptor antagonists, p38 MAPK inhibitors, caspase inhibitors and antioxidants (free-radical scavengers or other inhibitors of excessive nitric oxide or reactive oxygen species).

NMDAR antagonists have been shown to attenuate neuronal damage due to either HIV-infected macrophages or HIV/gp120, both *in vitro* and *in vivo*. Memantine, an open-channel blocker, prevents excessive NMDAR activity while sparing physiological function<sup>82,91</sup>. Also, unlike other NMDAR antagonists tested so far in clinical trials, memantine has proven both safe and effective in a number of phase III clinical trials for Alzheimer's disease and vascular dementia. The agent has been tested in patients with HAD in a large, multi-centre clinical trial sponsored by the US National Institutes of Health, the results of which are expected soon, and improved second-generation drugs are currently under development. Previous, small clinical trials of nimodipine, a voltage-activated calcium-channel blocker, and a PAF inhibitor suggested some therapeutic benefit but were not conclusive<sup>92,93</sup>. An additional clinical trial using the antioxidant drug selegiline is aimed at combating the effects of excitotoxicity by minimizing the impact of free radicals<sup>94</sup>.

Chemokine receptors allow HIV-1 to enter cells and as such are important potential therapeutic targets in the fight against AIDS<sup>45</sup>. Antagonists of CXCR4 and CCR5 inhibit HIV-1 entry and are being assessed in clinical trials<sup>45</sup>. However, the benefit of inhibitors of chemokine receptors for HIV-associated neurological complications, although likely, remains to be shown<sup>12</sup>. Interestingly, certain chemokines have been shown to protect neurons, even though the virus does not productively infect neurons. In particular,  $\beta$ -chemokines and fractalkine prevent gp120-induced neuronal apoptosis *in vitro*<sup>47,61,95</sup> and, similarly, some  $\beta$ -chemokines can ameliorate NMDAR-mediated neurotoxicity<sup>95</sup>. Distinct  $\beta$ -chemokines may therefore represent a potential treatment modality for HAD.

Neuronal apoptosis seems to be one of the hallmarks of neurodegenerative diseases including HAD<sup>32</sup>. Because caspases carry out the apoptotic programme, caspase inhibitors might be helpful in preventing detrimental neuronal loss<sup>96</sup>. Although caspase inhibitors are not currently available in a form deliverable to the CNS or targeted to degenerating neurons, with further advances in this field, such drugs might be developed. Care must be exerted to avoid inhibitors that promote oncogenic processes or interrupt physiological circuits.

Finally, p38 MAPK inhibitors have been shown to abrogate neuronal apoptosis due to excitotoxicity, HIV/gp120 exposure or  $\alpha$ -chemokine (SDF-1) toxicity. The pharmaceutical industry is currently developing p38 inhibitors for a variety of inflammatory- and stress-related conditions, such as arthritis, and this may expedite trials for CNS indications such as HAD. One important lesson from HAD is that the synergy between the excitatory and inflammatory pathways to neuronal damage may, at least in part, be common to other CNS disorders including stroke, spinal cord injury and Alzheimer's disease. As such, the development of new therapeutics for HAD is likely to have significant impact on several other important neurodegenerative diseases. □

- Dore, G. J. *et al.* Trends in incidence of AIDS illnesses in Australia from 1983 to 1994: the Australian AIDS cohort. *J. Acquir. Immune. Defic. Syndr. Hum. Retrovir.* **16**, 39–43 (1997).
- Ferrando, S. *et al.* Highly active antiretroviral treatment in HIV infection: benefits for neuropsychological function. *AIDS* **12**, F65–F70 (1998).
- Dore, G. J. *et al.* Changes to AIDS dementia complex in the era of highly active antiretroviral therapy. *AIDS* **13**, 1249–1253 (1999).
- Major, E. O., Rausch, D., Marra, C. & Clifford, D. HIV-associated dementia. *Science* **288**, 440–442 (2000).
- Clifford, D. B. Human immunodeficiency virus-associated dementia. *Arch. Neurol.* **57**, 321–324 (2000).
- Lipton, S. A. Treating AIDS dementia. *Science* **276**, 1629–1630 (1997).
- Lipton, S. A. & Gendelman, H. E. Dementia associated with the acquired immunodeficiency syndrome. *N. Engl. J. Med.* **332**, 934–940 (1995).
- McArthur, J. C. *et al.* Dementia in AIDS patients: incidence and risk factors. Multicenter AIDS Cohort Study. *Neurology* **43**, 2245–2252 (1993).
- Ellis, R. J. *et al.* Neurocognitive impairment is an independent risk factor for death in HIV infection. San Diego HIV Neurobehavioral Research Center Group. *Arch. Neurol.* **54**, 416–424 (1997).
- Sacktor, N. *et al.* HIV-associated neurologic disease incidence changes: Multicenter AIDS Cohort Study, 1990–1998. *Neurology* **56**, 257–260 (2001).
- Enting, R. H. *et al.* Antiretroviral drugs and the central nervous system. *AIDS* **12**, 1941–1955 (1998).
- Gartner, S. HIV infection and dementia. *Science* **287**, 602–604 (2000).
- Cunningham, P. H., Smith, D. G., Satchell, C., Cooper, D. A. & Brew, B. Evidence for independent development of resistance to HIV-1 reverse transcriptase inhibitors in the cerebrospinal fluid. *AIDS* **14**, 1949–1954 (2000).
- Asensio, V. C. & Campbell, I. L. Chemokines in the CNS: plurifunctional mediators in diverse states. *Trends. Neurosci.* **22**, 504–512 (1999).
- Prospero-Garcia, O. *et al.* Microglia-passaged simian immunodeficiency virus induces neurophysiological abnormalities in monkeys. *Proc. Natl Acad. Sci. USA* **93**, 14158–14163 (1996).
- Kalams, S. A. & Walker, B. D. Cytotoxic T lymphocytes and HIV-1 related neurologic disorders. *Curr. Top. Microbiol. Immunol.* **202**, 79–88 (1995).
- Sasseville, V. G. *et al.* Monocyte adhesion to endothelium in simian immunodeficiency virus-induced AIDS encephalitis is mediated by vascular cell adhesion molecule-1/alpha 4 beta 1 integrin interactions. *Am. J. Pathol.* **144**, 27–40 (1994).
- Fiala, M. *et al.* TNF- $\alpha$  opens a paracellular route for HIV-1 invasion across the blood-brain barrier. *Mol. Med.* **3**, 553–564 (1997).
- Toneatto, S., Finco, O., van der Putten, H., Abrignani, S. & Annunziata, P. Evidence of blood-brain barrier alteration and activation in HIV-1 gp120 transgenic mice. *AIDS* **13**, 2343–2348 (1999).
- Brew, B. J., Pemberton, L., Cunningham, P. & Law, M. G. Levels of human immunodeficiency virus type 1 RNA in cerebrospinal fluid correlate with AIDS dementia stage. *J. Infect. Dis.* **175**, 963–966 (1997).
- Ellis, R. J. *et al.* Cerebrospinal fluid human immunodeficiency virus type 1 RNA levels are elevated in neurocognitively impaired individuals with acquired immunodeficiency syndrome. HIV Neurobehavioral Research Center Group. *Ann. Neurol.* **42**, 679–688 (1997).
- McArthur, J. C. *et al.* Relationship between human immunodeficiency virus-associated dementia and viral load in cerebrospinal fluid and brain. *Ann. Neurol.* **42**, 689–698 (1997).
- Wiley, C. A. *et al.* Distribution of brain HIV load in AIDS. *Brain. Pathol.* **8**, 277–284 (1998).
- Glass, J. D., Fedor, H., Wesselingh, S. L. & McArthur, J. C. Immunocytochemical quantitation of human immunodeficiency virus in the brain: correlations with dementia. *Ann. Neurol.* **38**, 755–762 (1995).
- Maslah, E. *et al.* Dendritic injury is a pathological substrate for human immunodeficiency virus-related cognitive disorders. HNRC Group. The HIV Neurobehavioral Research Center. *Ann. Neurol.* **42**, 963–972 (1997).
- Wesselingh, S. L. *et al.* Cellular localization of tumor necrosis factor mRNA in neurological tissue from HIV-infected patients by combined reverse transcriptase/polymerase chain reaction in situ hybridization and immunohistochemistry. *J. Neuroimmunol.* **74**, 1–8 (1997).
- Giulian, D. *et al.* Study of receptor-mediated neurotoxins released by HIV-1-infected mononuclear phagocytes found in human brain. *J. Neurosci.* **16**, 3139–3153 (1996).
- Heyes, M. P. *et al.* Quinolinic acid in cerebrospinal fluid and serum in HIV-1 infection: relationship to clinical and neurological status. *Ann. Neurol.* **29**, 202–209 (1991).
- Everall, I. P. *et al.* Cortical synaptic density is reduced in mild to moderate human immunodeficiency virus neurocognitive disorder. HNRC Group. HIV Neurobehavioral Research Center. *Brain. Pathol.* **9**, 209–217 (1999).
- Fox, L., Alford, M., Achim, C., Mallory, M. & Maslah, E. Neurodegeneration of somatostatin-immunoreactive neurons in HIV encephalitis. *J. Neuropathol. Exp. Neurol.* **56**, 360–368 (1997).
- Maslah, E., Ge, N., Achim, C. L., Hansen, L. A. & Wiley, C. A. Selective neuronal vulnerability in HIV encephalitis. *J. Neuropathol. Exp. Neurol.* **51**, 585–593 (1992).
- Adle-Biasette, H. *et al.* Neuronal apoptosis in HIV infection in adults. *Neuropathol. Appl. Neurobiol.* **21**, 218–227 (1995).
- Gelbard, H. A. *et al.* Apoptotic neurons in brains from paediatric patients with HIV-1 encephalitis and progressive encephalopathy. *Neuropathol. Appl. Neurobiol.* **21**, 208–217 (1995).
- Petito, C. K. & Roberts, B. Evidence of apoptotic cell death in HIV encephalitis. *Am. J. Pathol.* **146**, 1121–1130 (1995).
- Adle-Biasette, H. *et al.* Neuronal apoptosis does not correlate with dementia in HIV infection but is related to microglial activation and axonal damage. *Neuropathol. Appl. Neurobiol.* **25**, 123–133 (1999).
- Giulian, D., Vaca, K. & Noonan, C. A. Secretion of neurotoxins by mononuclear phagocytes infected with HIV-1. *Science* **250**, 1593–1596 (1990).
- Giulian, D., Wendt, E., Vaca, K. & Noonan, C. A. The envelope glycoprotein of human immunodeficiency virus type 1 stimulates release of neurotoxins from monocytes. *Proc. Natl Acad. Sci. USA* **90**, 2769–2773 (1993).
- Lipton, S. A., Sucher, N. J., Kaiser, P. K. & Dreyer, E. B. Synergistic effects of HIV coat protein and NMDA receptor-mediated neurotoxicity. *Neuron* **7**, 111–118 (1991).
- Brew, B. J. *et al.* Quinolinic acid production is related to macrophage tropic isolates of HIV-1. *J. Neurovirol.* **1**, 369–374 (1995).
- Yeh, M. W. *et al.* Cytokine-stimulated, but not HIV-infected, human monocyte-derived macrophages produce neurotoxic levels of L-cysteine. *J. Immunol.* **164**, 4265–4270 (2000).
- Bonfoco, E., Krainc, D., Ankarcrona, M., Nicotera, P. & Lipton, S. A. Apoptosis and necrosis: two distinct events induced, respectively, by mild and intense insults with N-methyl-D-aspartate or nitric oxide/superoxide in cortical cell cultures. *Proc. Natl Acad. Sci. USA* **92**, 7162–7166 (1995).

42. Meucci, O. *et al.* Chemokines regulate hippocampal neuronal signaling and gp120 neurotoxicity. *Proc. Natl Acad. Sci. USA* **95**, 14500–14505 (1998).
43. Liu, Y. *et al.* Uptake of HIV-1 Tat protein mediated by low-density lipoprotein receptor-related protein disrupts the neuronal metabolic balance of the receptor ligands. *Nature Med.* **6**, 1380–1387 (2000).
44. He, J. *et al.* CCR3 and CCR5 are co-receptors for HIV-1 infection of microglia. *Nature* **385**, 645–649 (1997).
45. Michael, N. L. & Moore, J. P. HIV-1 entry inhibitors: evading the issue. *Nature Med.* **5**, 740–742 (1999).
46. Rottman, J. B. *et al.* Cellular localization of the chemokine receptor CCR5. Correlation to cellular targets of HIV-1 infection. *Am. J. Pathol.* **151**, 1341–1351 (1997).
47. Kaul, M. & Lipton, S. A. Chemokines and activated macrophages in HIV gp120-induced neuronal apoptosis. *Proc. Natl Acad. Sci. USA* **96**, 8212–8216 (1999).
48. Zheng, J. *et al.* Intracellular CXCR4 signaling, neuronal apoptosis and neuropathogenic mechanisms of HIV-1-associated dementia. *J. Neuroimmunol.* **98**, 185–200 (1999).
49. Sanders, V. J. *et al.* Chemokines and receptors in HIV encephalitis. *AIDS* **12**, 1021–1026 (1998).
50. McManus, C. M. *et al.* Chemokine and chemokine-receptor expression in human glial elements: induction by the HIV protein, Tat, and chemokine autoregulation. *Am. J. Pathol.* **156**, 1441–1453 (2000).
51. Zhang, L. *et al.* In vivo distribution of the human immunodeficiency virus/simian immunodeficiency virus coreceptors: CXCR4, CCR3, and CCR5. *J. Virol.* **72**, 5035–5045 (1998).
52. Limatola, C. *et al.* SDF-1 $\alpha$ -mediated modulation of synaptic transmission in rat cerebellum. *Eur. J. Neurosci.* **12**, 2497–2504 (2000).
53. Lazarini, F. *et al.* Differential signalling of the chemokine receptor CXCR4 by stromal cell- derived factor 1 and the HIV glycoprotein in rat neurons and astrocytes. *Eur. J. Neurosci.* **12**, 117–125 (2000).
54. Ohagen, A. *et al.* Apoptosis induced by infection of primary brain cultures with diverse human immunodeficiency virus type 1 isolates: evidence for a role of the envelope. *J. Virol.* **73**, 897–906 (1999).
55. Hesselgeser, J. *et al.* Neuronal apoptosis induced by HIV-1 gp120 and the chemokine SDF-1 $\alpha$  is mediated by the chemokine receptor CXCR4. *Curr. Biol.* **8**, 595–598 (1998).
56. Letendre, S. L., Lanier, E. R. & McCutchan, J. A. Cerebrospinal fluid beta chemokine concentrations in neurocognitively impaired individuals infected with human immunodeficiency virus type 1. *J. Infect. Dis.* **180**, 310–319 (1999).
57. Cinque, P. *et al.* Elevated cerebrospinal fluid levels of monocyte chemoattractant protein-1 correlate with HIV-1 encephalitis and local viral replication. *AIDS* **12**, 1327–1332 (1998).
58. Kelder, W., McArthur, J. C., Nance-Sproson, T., McClernon, D. & Griffin, D. E. Beta-chemokines MCP-1 and RANTES are selectively increased in cerebrospinal fluid of patients with human immunodeficiency virus-associated dementia. *Ann. Neurol.* **44**, 831–835 (1998).
59. Combadiere, C. *et al.* Identification of CX3CR1. A chemotactic receptor for the human CX3C chemokine fractalkine and a fusion coreceptor for HIV-1. *J. Biol. Chem.* **273**, 23799–23804 (1998).
60. Tong, N. *et al.* Neuronal fractalkine expression in HIV-1 encephalitis: roles for macrophage recruitment and neuroprotection in the central nervous system. *J. Immunol.* **164**, 1333–1339 (2000).
61. Meucci, O., Fatatis, A., Simen, A. A. & Miller, R. J. Expression of CX3CR1 chemokine receptors on neurons and their role in neuronal survival. *Proc. Natl Acad. Sci. USA* **97**, 8075–8080 (2000).
62. Maciejewski-Lenoir, D., Chen, S., Feng, L., Maki, R. & Bacon, K. B. Characterization of fractalkine in rat brain cells: migratory and activation signals for CX3CR1-expressing microglia. *J. Immunol.* **163**, 1628–1635 (1999).
63. Chapman, G. A. *et al.* Fractalkine cleavage from neuronal membranes represents an acute event in the inflammatory response to excitotoxic brain damage. *J. Neurosci.* **20**, RC87 (2000).
64. Canki, M. *et al.* Isolation and long-term culture of primary ocular human immunodeficiency virus type 1 isolates in primary astrocytes. *J. Neurovirol.* **3**, 10–15 (1997).
65. Dreyer, E. B., Kaiser, P. K., Offermann, J. T. & Lipton, S. A. HIV-1 coat protein neurotoxicity prevented by calcium channel antagonists. *Science* **248**, 364–367 (1990).
66. Bhat, N. R., Zhang, P., Lee, J. C. & Hogan, E. L. Extracellular signal-regulated kinase and p38 subgroups of mitogen-activated protein kinases regulate inducible nitric oxide synthase and tumor necrosis factor- $\alpha$  gene expression in endotoxin-stimulated primary glial cultures. *J. Neurosci.* **18**, 1633–1641 (1998).
67. Toggas, S. M. *et al.* Central nervous system damage produced by expression of the HIV-1 coat protein gp120 in transgenic mice. *Nature* **367**, 188–193 (1994).
68. Wyss-Coray, T. *et al.* Dysregulation of signal transduction pathways as a potential mechanism of nervous system alterations in HIV-1 gp120 transgenic mice and humans with HIV-1 encephalitis. *J. Clin. Invest.* **97**, 789–798 (1996).
69. Fine, S. M. *et al.* Tumor necrosis factor  $\alpha$  inhibits glutamate uptake by primary human astrocytes. Implications for pathogenesis of HIV-1 dementia. *J. Biol. Chem.* **271**, 15303–15306 (1996).
70. Genis, P. *et al.* Cytokines and arachidonic metabolites produced during human immunodeficiency virus (HIV)-infected macrophage-astroglia interactions: implications for the neuropathogenesis of HIV disease. *J. Exp. Med.* **176**, 1703–1718 (1992).
71. Bezzi, P. *et al.* Prostaglandins stimulate calcium-dependent glutamate release in astrocytes. *Nature* **391**, 281–285 (1998).
72. Conant, K. *et al.* Induction of monocyte chemoattractant protein-1 in HIV-1 Tat-stimulated astrocytes and elevation in AIDS dementia. *Proc. Natl Acad. Sci. USA* **95**, 3117–3121 (1998).
73. Adamson, D. C. *et al.* Immunologic NO synthase: elevation in severe AIDS dementia and induction by HIV-1 gp41. *Science* **274**, 1917–1921 (1996).
74. Hesselgeser, J. *et al.* CD4-independent association between HIV-1 gp120 and CXCR4: functional chemokine receptors are expressed in human neurons. *Curr. Biol.* **7**, 112–121 (1997).
75. Fontana, G., Valenti, L. & Raiteri, M. gp120 can revert antagonism at the glycine site of NMDA receptors mediating GABA release from cultured hippocampal neurons. *J. Neurosci. Res.* **49**, 732–738 (1997).
76. Nath, A. *et al.* Synergistic neurotoxicity by human immunodeficiency virus proteins Tat and gp120: protection by memantine. *Ann. Neurol.* **47**, 186–194 (2000).
77. Pillar, S. C., Jans, P., Gage, P. W. & Jans, D. A. Extracellular HIV-1 virus protein R causes a large inward current and cell death in cultured hippocampal neurons: implications for AIDS pathology. *Proc. Natl Acad. Sci. USA* **95**, 4595–4600 (1998).
78. Shi, B. *et al.* Apoptosis induced by HIV-1 infection of the central nervous system. *J. Clin. Invest.* **98**, 1979–1990 (1996).
79. Xiong, H. *et al.* HIV-1 infected mononuclear phagocyte secretory products affect neuronal physiology leading to cellular demise: relevance for HIV-1-associated dementia. *J. Neurovirol.* **6**(Suppl. 1), S14–S23 (2000).
80. Pulliam, L., Herndird, B. G., Tang, N. M. & McGrath, M. S. Human immunodeficiency virus-infected macrophages produce soluble factors that cause histological and neurochemical alterations in cultured human brains. *J. Clin. Invest.* **87**, 503–512 (1991).
81. Brennenman, D. E. *et al.* Neuronal cell killing by the envelope protein of HIV and its prevention by vasoactive intestinal peptide. *Nature* **335**, 639–642 (1988).
82. Lipton, S. A. Prospects for clinically tolerated NMDA antagonists: open-channel blockers and alternative redox states of nitric oxide. *Trends. Neurosci.* **16**, 527–532 (1993).
83. Gendelman, H. E., Genis, P., Jett, M., Zhai, Q. H. & Nottet, H. S. An experimental model system for HIV-1-induced brain injury. *Adv. Neuroimmunol.* **4**, 189–193 (1994).
84. Toggas, S. M., Masliah, E. & Mucke, L. Prevention of HIV-1 gp120-induced neuronal damage in the central nervous system of transgenic mice by the NMDA receptor antagonist memantine. *Brain. Res.* **706**, 303–307 (1996).
85. Budd, S. L., Tannetti, L., Lishnak, T. & Lipton, S. A. Mitochondrial and extramitochondrial apoptotic signaling pathways in cerebrocortical neurons. *Proc. Natl Acad. Sci. USA* **97**, 6161–6166 (2000).
86. Tannetti, L., D'Emilia, D. M., Troy, C. M. & Lipton, S. A. Role of caspases in N-methyl-D-aspartate-induced apoptosis in cerebrocortical neurons. *J. Neurochem.* **71**, 946–959 (1998).
87. Ghatan, S. *et al.* p38 MAP kinase mediates bax translocation in nitric oxide-induced apoptosis in neurons. *J. Cell Biol.* **150**, 335–347 (2000).
88. Garden, G. A., Budd, S. L., Kaul, M. & Lipton, S. A. HIV-gp120 induces apoptosis in cultured cortical neurons by pathways involving caspases 8 and 9 and TNF- $\alpha$ . *Soc. Neurosci. Abstr.* **26**, 1062 (2000).
89. Pulliam, L., Zhou, M., Stubblebine, M. & Bitler, C. M. Differential modulation of cell death proteins in human brain cells by tumor necrosis factor alpha and platelet activating factor. *J. Neurosci. Res.* **54**, 530–538 (1998).
90. Shi, B., Raina, J., Lorenzo, A., Busciglio, J. & Gabuzda, D. Neuronal apoptosis induced by HIV-1 Tat protein and TNF- $\alpha$ : potentiation of neurotoxicity mediated by oxidative stress and implications for HIV-1 dementia. *J. Neurovirol.* **4**, 281–290 (1998).
91. Lipton, S. A. Memantine prevents HIV coat protein-induced neuronal injury in vitro. *Neurology* **42**, 1403–1405 (1992).
92. Schifitto, G. *et al.* Randomized trial of the platelet-activating factor antagonist lexipafant in HIV-associated cognitive impairment. Neurological AIDS Research Consortium. *Neurology* **53**, 391–396 (1999).
93. Navia, B. A. *et al.* A phase I/II trial of nimodipine for HIV-related neurologic complications. *Neurology* **51**, 221–228 (1998).
94. Sacktor, N. *et al.* Transdermal selegiline in HIV-associated cognitive impairment: pilot, placebo-controlled study. *Neurology* **54**, 233–235 (2000).
95. Bruno, V., Copani, A., Besong, G., Scoto, G. & Nicoletti, F. Neuroprotective activity of chemokines against N-methyl-D-aspartate or  $\beta$ -amyloid-induced toxicity in culture. *Eur. J. Pharmacol.* **399**, 117–121 (2000).
96. Yuan, J. & Yankner, B. A. Apoptosis in the nervous system. *Nature* **407**, 802–809 (2000).

# Acknowledgements

We dedicate this work to the memory of T. Sandson for his compassionate treatment of patients with HIV-associated dementia. We sincerely apologize to our colleagues whose works we could not cite owing to space and reference limitations. We thank members of our laboratory for their hard work and helpful discussions. M.K. is supported by the American Foundation for AIDS Research. S.A.L. and G.A.G. are supported by the National Institutes of Health. S.A.L. is or has recently been a consultant to Allergan, Alcon and Neurobiological Technologies, Inc. in the field of neuroprotective agents.



# HIV chemotherapy

Douglas D. Richman

Veterans Affairs San Diego Healthcare System and University of California San Diego, Departments of Pathology and Medicine 0679, 9500 Gilman Drive, La Jolla, California 92093-0679, USA (e-mail: drichman@ucsd.edu)

**The use of chemotherapy to suppress replication of the human immunodeficiency virus (HIV) has transformed the face of AIDS in the developed world. Pronounced reductions in illness and death have been achieved and healthcare utilization has diminished. HIV therapy has also provided many new insights into the pathogenesis and the viral and cellular dynamics of HIV infection. But challenges remain. Treatment does not suppress HIV replication in all patients, and the emergence of drug-resistant virus hinders subsequent treatment. Chronic therapy can also result in toxicity. These challenges prompt the search for new drugs and new therapeutic strategies to control chronic viral replication.**

**H**IV infection is characterized by a prolonged asymptomatic period of years to decades, which is followed by the fatal illness of AIDS. Various complications characterize AIDS, including wasting, neurological impairment, and opportunistic infections and malignancies. The asymptomatic period was often considered as relatively quiescent with regard to viral replication with the frequent usage of the misnomer 'clinical latency'. But several studies in the 1980s and early 1990s documented evidence of extensive replication during this time (reviewed in ref. 1).

The development of quantitative assays for HIV RNA in plasma and the use of antiviral drugs to perturb quasi steady-state levels of this RNA facilitated an understanding of HIV dynamics<sup>2-4</sup>. Several months after infection, a relatively stable level of between  $10^3$  and  $10^6$  copies of HIV RNA per millilitre of plasma is achieved in each individual. This level of steady-state RNA, which is in the form of circulating virus particles, is determined primarily by host genetics, although the virulence of an individual virus isolate may also contribute. Genetic polymorphisms in chemokine receptors and haplotypes of the human leukocyte antigen are the best-defined host determinants<sup>5,6</sup>.

The administration of antiretroviral therapy results in the rapid reduction of HIV RNA levels (Fig. 1a). Assuming therapy is inhibiting virus production almost completely, calculations can then be made of rates of virus clearance and production (which must be equivalent in a steady state)<sup>3,4,7</sup>. HIV virions (mature, infectious virus particles) have a half-life in the plasma of less than 6 hours<sup>7</sup>. Infected activated CD4 lymphocytes produce at least 99% of circulating virus and these cells have a half-life of 1.6 days. At least  $10^{10}$  virions are generated daily and approximately 140 generational cycles occur annually<sup>7</sup>. Thus, even during asymptomatic infection the magnitude and dynamics of virus infection with which chemotherapy (and the immune system) must contend is formidable.

## Pathogenesis and natural history

Approximately  $10^7$ – $10^8$  CD4 lymphocytes (about 0.1% of the body's total) are productively infected with HIV at any given time and perhaps 100 times that number contain HIV nucleic acid<sup>8,9</sup>. The rates of virus clearance seem similar in all patients and all disease stages, which implies that steady-state levels of HIV RNA in the blood are determined by the rate of virus production. These rates of production are a function of the number of infected lymphocytes in the lymphoid tissue<sup>8,9</sup>, so that steady-state levels of HIV RNA are

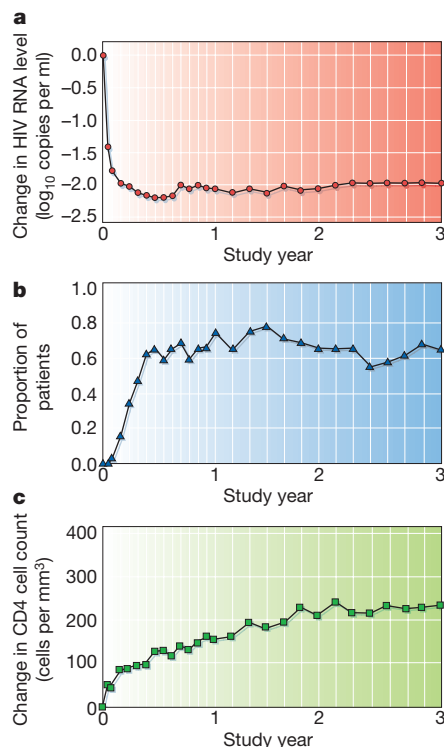
related directly to the rate of decline of CD4 lymphocytes. The higher the RNA levels the faster the loss of CD4 cells and the shorter the duration of HIV infection before death<sup>10</sup>. Because the CD4 count determines the risk of disease and death and the level of HIV RNA determines the rate of CD4 cell decline, these values are used routinely to assess clinical status and timescales for initiation of chemotherapy<sup>11</sup>.

The pathogenesis of HIV and its treatment are complicated by the existence of tissue compartments and cellular reservoirs. Gut-associated and nodal lymphoid tissues seem to be infected at different rates during acute infection with simian immunodeficiency virus (SIV) and to be differentially susceptible to virus that uses different chemokine receptors<sup>12,13</sup>. Although there is movement between the blood and central nervous system (CNS), much of the virus in the CNS evolves independently<sup>14-16</sup>. Similar observations have been made with virus in semen<sup>17,18</sup>.

Activated CD4 lymphocytes and macrophages have long been appreciated as the predominant host cells for HIV replication. The inhibition of most virus replication with potent chemotherapy exposed the latently infected CD4 lymphocyte<sup>19-21</sup>. These cells represent a small fraction of infected cells during active infection and have a half-life of at least 6 months<sup>22,23</sup>. Consequently, many such cells survive for years, archiving virus that can re-emerge and propagate after the withdrawal of chemotherapy. The existence of this cellular reservoir has frustrated hopes of eradication of infection with current chemotherapy.

## Antiretroviral therapy Impact on viral dynamics

When potent combination therapy is administered effectively, levels of HIV RNA in plasma and infected cells in lymphoid tissue clear rapidly (Fig. 1a). This first-phase clearance is attributable to the death of infected activated CD4 lymphocytes and the prevention of new infections<sup>7</sup>. The second-phase clearance rate is slower and somewhat more variable in slope among individuals. It has been attributed to clearance of infected macrophages or to virions bound to dendritic cells in lymph nodes, but may also be due to chronically infected CD4 lymphocytes with a longer half-life and lower rates of virus replication<sup>24-26</sup>. Failure to reduce HIV RNA to below 50 copies per millilitre plasma (the limit of detection in currently available assays) indicates inadequate suppression and risk of outgrowth of resistant virus<sup>11</sup>. Many patients sustaining suppression below this 50-copy value (Fig. 1b) have done so for more than 5 years. But examination of tissues and blood for viral RNA and for closed circular reverse transcripts, which



**Figure 1** Three-year treatment with indinavir, zidovudine and lamivudine. **a**, Median changes in serum HIV RNA level (Amplicor assay with quantification limit of 500 copies per millilitre) from baseline. **b**, Proportions of patients with serum HIV RNA levels less than 50 copies per millilitre (ultradirect assay). **c**, Median changes in CD4 cell count from baseline. The number of contributing patients in each trial and at each time point was between 30 and 33. Details regarding the study and analyses are published<sup>33,106</sup>. (Adapted from ref. 106 with permission.)

have a short half-life, as well as documentation of nucleotide sequence evolution, all indicate that most effectively treated patients are not completely suppressing virus replication but instead are experiencing intermittent or smouldering virus replication that is difficult to discern<sup>23,26–32</sup>.

#### Restoration of immune function

The immunological consequences of suppressing virus replication are pronounced (Fig. 1c). The increase in CD4 lymphocyte numbers has two phases. In the first month or two the increase is often large (20–100 cells per  $\mu$ l blood)<sup>33–36</sup>. The magnitude is proportional to the steady-state HIV RNA levels, which correspond to the level of generalized activation of the immune system. The normal distribution of lymphocytes is 2% in the circulation and 98% in the lymphoid tissues. With the immune activation of HIV infection the ratio shifts to 1% and 99%, respectively<sup>9,37</sup>. Therapy largely corrects this shift and results in redistribution of mostly CD45RO<sup>+</sup> memory T cells from the lymphoid tissue back to the circulation<sup>35,37,38</sup>. Production of new cells, mostly of the CD45RA naive phenotype, is generated both by peripheral proliferation and by restored thymic mass and function, which is age related<sup>39,40</sup>.

But generation of cell numbers is not sufficient. It is the restoration of immune function that has transformed the natural history of AIDS. Both CD4 and CD8 responses to recall antigens are regenerated<sup>41–43</sup>, and persistent opportunistic infections are often resolved. Occasionally, subclinical chronic infections are manifested when a restored immune response produces a local inflammatory reaction<sup>44–46</sup>. Patient care has been transformed with the ability to withdraw prophylactic or suppressive chemotherapy for pneumocystis, toxoplasmosis, cytomegalovirus, *Mycobacterium avium* complex, leishmaniasis, cryptococcosis and candidal thrush, which previously had been lifelong afflictions<sup>47</sup>.

#### Impact on natural history of AIDS

AIDS mortality statistics reflect drug usage in Western Europe and the Americas (Fig. 2)<sup>48–51</sup>. Opportunistic disease has been reversed and prevented, healthcare costs have diminished, and many ill and disabled patients have returned to normal and functional life styles. But the dramatic impact of restoration of immune function does come with a cost — the expense, inconvenience and toxicity of antiretroviral therapy.

#### Antiretroviral drugs

The genetic efficiency of small viruses indicates that any viral target could be a candidate for chemotherapeutic inhibition. The drug discovery process usually requires the design of high-throughput screening assays with targets whose genetic deletion renders the virus non-infectious. It is for this reason that the enzymes reverse transcriptase and protease became the first viral targets (Table 1). Nucleoside analogues were already being used as polymerase inhibitors to treat malignancies and herpesvirus infections, and the screening of nucleosides as inhibitors of reverse transcriptase led to the first class of antiretroviral drugs. The nucleoside inhibitors of reverse transcriptase all undergo anabolic phosphorylation by host-cell enzymes to generate the 5'-triphosphate. All antiretroviral nucleosides lack 3'-OH moieties, thus preventing the formation of 3'-5' phosphodiester bonds between the elongating DNA chain and incoming 5'-nucleoside triphosphates. The antiretroviral nucleosides therefore act as chain terminators when incorporated into reverse transcripts (Fig. 3).

The non-nucleoside inhibitors of reverse transcriptase represent numerous chemically divergent, polycyclic compounds that all bind in a site near the polymerase catalytic domain<sup>52</sup>. It is of note that the crystal structure of reverse transcriptase would not have aided drug discovery as the binding pocket of these drugs does not exist in the enzyme in the absence of bound drug<sup>53</sup>.

The protease inhibitors were designed as peptidomimetics of the viral peptide substrates with non-cleavable structures in the scissile bond (Fig. 4). The prolonged development time between the discovery of inhibitors and the development of drugs resulted from the challenge of modifying potent inhibitory compounds into drugs that were soluble and orally bioavailable.

#### Challenges to effective therapy

##### Drug potency

Each of the approved antiviral drugs can reduce plasma HIV RNA by 0.5–2.0 logarithmic<sub>10</sub> units, with a corresponding improvement in CD4 cell counts. But only by combining three or more drugs has the potency been sufficient to generate sustained suppression without outgrowth of resistant virus (Fig. 1b)<sup>33</sup>. However, because of problems with adherence, pharmacology and toxicity, only 50–90% of study subjects achieve the desired suppression with current regimens (Fig. 1b). Moreover, as mentioned above, most patients with sustained suppression still exhibit low levels of replication.

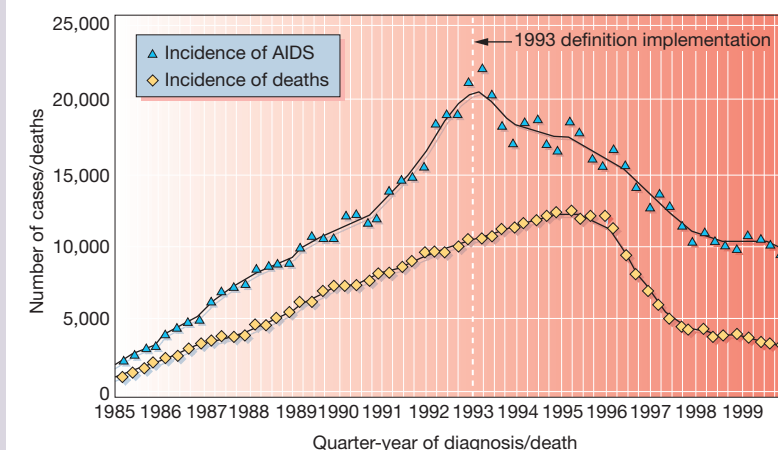
##### Adherence

Zidovudine (3'-azido-3'-deoxythymidine or AZT) initially had to be ingested every 4 hours, and combinations required administration three times daily with adjustment for meals. Most regimens are now taken twice daily, but they must be taken conscientiously on a daily basis for years. Numerous studies have documented a correlation between successful suppression and adherence to medication schedules<sup>34,55</sup>. Education and support by healthcare providers, coupled with commitment by patients despite side effects and inconvenience, has become a critical component of successful treatment.

##### Pharmacology

Variable drug absorption and drug interactions plague combination therapy. Inter-individual variability in plasma drug levels after the same dose of drug contributes to drug failure at one extreme and toxicity at the other<sup>56,57</sup>. Emerging data indicate that host genetic polymorphisms

**Figure 2** Estimated incidence of AIDS and deaths of adults/adolescents with AIDS in the United States during the period 1985–1999. Number of deaths is adjusted for reporting delays. (Reproduced from ref. 107 with permission.)



may explain much of this variability. Interactions with other drugs (prescribed, recreational and alternative) and diet add to this variability. Potential solutions include the design of drugs with less inter-individual pharmacological variability or with greater therapeutic indices to tolerate greater margins of error, monitoring drug levels, or genotyping patients to select drugs or drug doses. Most non-nucleoside reverse-transcriptase inhibitors and protease inhibitors are inducers, substrates or inhibitors to various degrees of the hepatic cytochrome P450 system. The consequences of interactions are complicated, and unpredictable changes in drug levels to subtherapeutic or toxic concentrations can occur when drugs are combined with each other or with numerous other drugs that interact with these enzymes<sup>58</sup>.

This interaction has been exploited for protease inhibitor therapy. Ritonavir is both difficult to tolerate in full doses and a potent inhibitor of cytochrome P450 3A4 isoform (CYP3A4), a hepatic enzyme that detoxifies xenobiotics (foreign substances such as drugs). At small doses it is now used routinely with most other protease inhibitors to improve their pharmacokinetic profile and provide more effective drug levels and less frequent dosing.

#### Tissue compartments

Impaired distribution of drugs into tissue compartments such as the brain and genital tract could result in local replication of virus. Some compounds penetrate these pharmacological sanctuaries poorly<sup>59</sup>, and certain protease inhibitors are actively transported out of the CNS by a multidrug-resistant P-glycoprotein<sup>60</sup>.

#### Toxicity

No drug is without toxicity. When antiretroviral drugs were first introduced, risks and toxicities were tolerated in the face of imminently life-threatening disease. With prolonged chronic therapy and the disappearance of the common symptoms of underlying HIV disease, adverse complications of antiretroviral drugs are being identified and characterized. As patients assume asymptomatic lives with hopes of prolonged survival, several toxicities represent increasing concerns regarding antiretroviral drug use. Protease inhibitors have been associated with insulin intolerance, elevation of cholesterol and triglyceride concentrations, and redistribution of body fat, with loss of fat from face and limbs and central adiposity<sup>61–63</sup>. New challenges include elucidating toxic mechanisms, managing these side effects and designing drugs to segregate antiviral and toxic activities.

### HIV drug resistance resulting from incomplete suppression

#### The biology of resistance

Approximately 10 billion ( $10^{10}$ ) HIV-1 virions are generated daily<sup>7</sup>. With a mutation rate of about  $10^{-5}$  nucleotides per replication cycle and no proofreading mechanism for reverse transcription, approximately one mutation is generated for each new genome of 9,200

nucleotides<sup>64</sup>. Thus genomes with each possible mutation as well as many with double mutations should be generated daily. Incompletely suppressed viral replication, on a scale sufficient to exert selective pressure, drives the evolution and fixation of drug-resistant virus at a rate Darwin himself never imagined. Moreover, drug-resistant virus is readily archived in latently infected cells to confound treatment modifications for the remainder of the patient's life<sup>19,20</sup>.

#### Impact of resistance on treatment

As resistance mutations accumulate, drug susceptibility diminishes, progressively reducing the potency of components of combination regimens. Continued replication in the presence of drug selects for even greater levels of resistance to each administered drug and progressive cross-resistance to drugs of the same class. Thus, although impotent regimens, suboptimal adherence, pharmacological hurdles and ineffectively treated compartments may permit the emergence of resistant virus, its emergence drives a vicious cycle of treatment failure and yet more difficult treatment challenges. New regimens devised for patients failing treatment will not only be constrained by more limited options resulting from resistance, but also must still contend with the same obstacles that challenged the first regimen.

#### Transmission

Resistant virus in blood or genital secretions can be transmitted during sex, needle sharing or childbirth<sup>65</sup>. Documentation of the transmission of drug resistance has expanded from anecdotes of transmission of AZT-resistant virus to the identification of large cohorts with 10% or more of new infections due to drug-resistant virus, half of which exhibit resistance to multiple classes of drug<sup>65,66</sup>. Such patients are more likely to fail their first treatment regimen.

#### Resistance testing to manage patients

The accumulation of drug resistance due to treatment failure and transmission raises challenges to the effective treatment of individuals and to public health similar to those resulting from widespread

**Table 1** Approved antiretroviral drugs

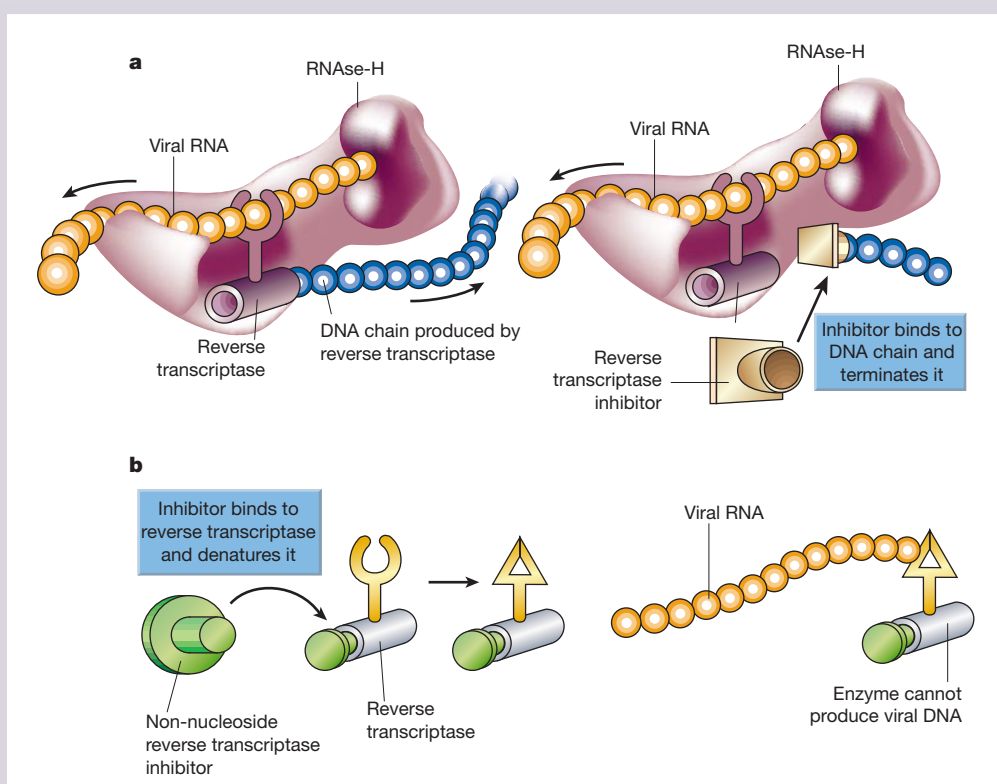
|                                                 |                   |
|-------------------------------------------------|-------------------|
| Nucleoside reverse-transcriptase inhibitors     |                   |
| Abacavir (ABC)                                  | Stavudine (d4T)   |
| Didanosine (ddI)                                | Zalcitabine (ddC) |
| Lamivudine (3TC)                                | Zidovudine (AZT)  |
| Non-nucleoside reverse-transcriptase inhibitors |                   |
| Efavirenz                                       | Nevirapine        |
| Delavirdine                                     |                   |
| Protease inhibitors                             |                   |
| Amprenavir*                                     | Nelfinavir        |
| Indinavir*                                      | Ritonavir         |
| Lopinavir*                                      | Saquinavir*       |

\*Often or usually used with low-dose zidovudine for pharmacological enhancement.



**Figure 3** Mechanism of action of nucleoside and non-nucleoside reverse-transcriptase inhibitors. To enable HIV to be integrated into the host DNA and so use the cell's genetic machinery to make new virus, the single-stranded viral RNA must first be converted to double-stranded DNA by the viral enzyme reverse transcriptase, while the enzyme RNase-H hydrolyses the RNA after it has been copied. Nucleoside and non-nucleoside reverse-transcriptase inhibitors are two classes of antiretroviral drugs that suppress HIV replication by attacking reverse transcriptase.

**a**, Nucleoside reverse-transcriptase inhibitors are similar in structure to the building blocks that make up DNA. By incorporating themselves into the DNA nucleoside chain being produced by reverse transcriptase, they stop attachment of further nucleosides and so prevent ongoing viral DNA synthesis. **b**, Non-nucleoside reverse transcriptase inhibitors attach to the reverse transcriptase and affect the activity of the enzyme by restricting its mobility and making it unable to function. (Adapted from ref. 108 with permission.)



antibiotic resistance. Testing for drug resistance is rapidly being incorporated into standard HIV care<sup>67,68</sup>. Assays can help determine which drugs will not work (thereby diminishing cost, toxicity and inconvenience) and which drugs are most likely to be effective.

Two types of drug-resistance assays are being used for patient management. Phenotypic assays that measure drug susceptibility use a portion of a *gag-pol* fragment, amplified by polymerase chain reaction, from plasma RNA. This is incorporated into a recombinant vector plasmid that is used in high-throughput assays using reporter cells in various concentrations of drug<sup>69,70</sup>. Most genotypic assays are now performing dideoxynucleotide terminator-cycle sequencing of these same amplicons generated from HIV plasma RNA. Interpretation of the results of genotypic assays is based on presumed susceptibilities inferred from the combinations of mutations identified from the growing list of mutations known to confer phenotypic changes<sup>67,68</sup>.

Retrospective studies have shown consistently that both phenotypic and genotypic drug-resistance assays predict response to treatment<sup>71</sup>. Several prospective trials have shown that the treatment of patients failing a chemotherapeutic regimen is more successful with information from drug-resistance testing<sup>72,73</sup>.

## Improving the efficacy of treatment

### Better use of existing drugs

With the potency of current combination therapy at the threshold of suppression there is little margin for error. Suboptimal prescribing patterns by physicians who fail to keep pace with this highly complex and rapidly changing field, and suboptimal adherence by patients faced with the daily task of administering several doses of many pills, both further compromise efficacy. The challenges of modifying human behaviour and of adhering to difficult regimens for years dictate the need to develop more effective and forgiving drugs.

### Designing new and better drugs

**Toxicity.** Toxicity impacts on both the patient's ability to sustain adherence to a chronic regimen and the willingness of physicians and patients to initiate regimens. The recognition of the effects of protease inhibitors has forced a shift in prescribing patterns towards

non-nucleoside reverse-transcriptase inhibitors and towards delaying the initiation of treatment until more advanced stages of disease. Ignorance of the mechanisms underlying the metabolic toxicities is a barrier to addressing them effectively. The presumption that the mechanisms of antiviral activity and toxicity are non-identical provides the hope that drug design can better segregate the two and select compounds with superior therapeutic indices.

**Pharmacokinetics.** When the half-life of effective drug levels is prolonged, the frequency of drug administration can be reduced, creating a buffer for sustaining effective drug levels when patients are intermittently nonadherent. Some reverse-transcriptase inhibitors can be administered once daily and the use of low-dose zidovudine to inhibit CYP3A4 and thus prolong the half-lives of other protease inhibitors already provides a step towards once-daily and more forgiving regimens. The challenge is to develop new drugs that will permit more alternatives for once-daily regimens that are both compatible pharmacologically and effective virologically.

The ritonavir-enhanced protease-inhibitor regimens also can increase plasma levels of drugs. These higher drug levels can be sufficient to inhibit virus with low or moderate levels of resistance that would not be amenable to effective inhibition with the non-enhanced plasma levels<sup>74,75</sup>.

**Resistance profiles.** The accumulation of resistance mutations with failing regimens tends to select for progressively increasing cross-resistance within each of the three classes of available drugs. One approach to the treatment of resistance is pharmacologically enhanced drug levels of available drugs, as mentioned above. A second approach is the design of compounds with greater potency to compensate for the loss of activity against relatively resistant viruses. A third approach is the design of compounds with activity against resistant virus. Some compounds with these characteristics are in clinical trial, and others designed with activity against resistant enzymes are being developed<sup>76</sup>.

**New targets.** Resistance to protease or reverse-transcriptase inhibitors should not affect compounds designed against new targets. Although any function in a genetically efficient organism is a candidate for an inhibitory drug, potent compounds have been reported against only three additional targets.

Compounds targeting the zinc fingers of *gag* — a gene coding for the HIV core protein — potentially inhibit virus infectivity<sup>77</sup>. Incorporating such activity into an orally bioavailable, clinically useful therapeutic is the next step. Integrase is a viral enzyme that mediates integration of the viral reverse transcripts into the host chromosome and is essential for virus replication. Potent inhibitors have been identified<sup>78</sup>. These select for resistance mutations in the target enzyme, providing clear evidence of specific antiviral activity, although so far no integrase inhibitors have entered clinical trials.

Inhibitors of HIV fusion have been identified that target both the virus surface glycoprotein and the host-cell chemokine receptors targeted by the virus<sup>79</sup>. T-20 and T-1249 are peptides that correspond to an extracellular domain of the transmembrane segment (gp41) of the HIV envelope glycoprotein. Host cell-receptor binding induces a conformational change in gp41 resulting in the 'spring-loaded' formation of coiled-coil helices that then mediates fusion of the viral envelope and host-cell membrane<sup>80</sup>. Clinical trials have been encouraging, although the potential for resistance and lack of oral bioavailability remain challenges. But documentation of 2log<sub>10</sub> reductions in plasma HIV RNA with 14 days of monotherapy has proved that fusion is an alternative target with clinical potential and has prompted the search for better inhibitors of this target<sup>81,82</sup>.

CXCR4 and CCR5, the two chemokine receptors used by HIV together with CD4 for its host-cell receptors, have been targeted with small molecules to inhibit HIV replication *in vitro*<sup>83,84</sup>. Clinical trials have been initiated to determine whether clinical activity can be documented. Concerns include the risk of targeting a host-cell function and the capacity of HIV to shift its use of chemokine co-receptor to escape chemotherapeutic inhibitors.

### Implications for the developing world

#### Treatment is for the economically privileged

Although the cost of HIV care has fallen slowly since the introduction of AZT, costs in the United States were still at least US\$20,000 per patient per year at the time when protease inhibitors were introduced in 1995<sup>85,86</sup>. With the advent of more potent and more expensive antiretroviral therapy, overall healthcare costs have dropped in proportion to the reductions in serious and costly complications of advanced HIV<sup>87,88</sup>.

Even though these therapies are cost-effective by western standards at US\$10,000–15,000 per year of life saved, most developing

countries cannot afford the drugs as well as the necessary monitoring and support<sup>89,90</sup>. But economic barriers to use are not confined to the developing world. Even in the United States, where reductions of 80% in mortality rates have been documented among those receiving optimal treatment, persons of low socioeconomic status or with emotional and behavioural impairments have difficulty accessing appropriate therapy<sup>91</sup>.

#### Potential applications in the developing world

The resources do not exist to implement chronic chemotherapy in many developing countries even if the drugs were to be provided at low cost. Moreover, the complexity of antiretroviral chemotherapy, requiring expertise and costly high-technology laboratory capabilities, raise daunting challenges. The use of antiretroviral drugs in the peripartum period to reduce rates of maternal–fetal transmission has been demonstrated to be a cost-effective and feasible use of antiretroviral drugs in developing countries<sup>92,93</sup>. Some locations are documenting HIV seroprevalence in almost one-half of pregnant women with rates of transmission to children exceeding 30%.

### Therapeutic immunization

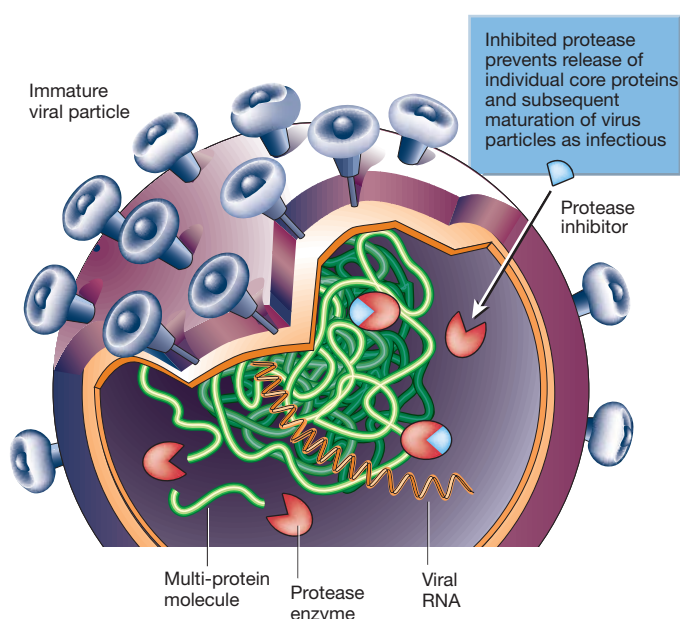
#### HIV immunity and natural history

Stronger immune responses correlate with less aggressive disease course. CD4 cell-proliferative responses to HIV antigen, CD8 cytotoxic T-cell responses and neutralizing antibody levels are all higher in long-term non-progressors than in patients with progressive disease<sup>94–98</sup>, raising the question about causal relationships. Diminution of virus replication after acute infection is associated temporally with the appearance of HIV-specific cytotoxic T-lymphocyte effector responses<sup>99</sup>. Perhaps more compelling are studies with rhesus macaques infected with SIV and SHIV (a genetically engineered hybrid virus having an HIV envelope and an SIV core) that support the contention that CD8 T cells, by mechanisms not yet elucidated, are required for control of HIV replication. With this model, the *in vivo* depletion of CD8 cells by monoclonal antibodies was associated temporally with significantly higher levels of viral replication<sup>100–102</sup>. With the recovery of CD8 cells, the SIV levels diminished<sup>101,102</sup>.

#### HIV-specific therapeutic immunization

With accumulating evidence that HIV (and SIV)-specific immunity is protective and with evidence that chemotherapy can result in

**Figure 4** Mechanism of action of protease inhibitors. After transcription in the nucleus, viral mRNA enters the cytoplasm and uses the host's cellular machinery to manufacture virus proteins. The viral components then gather at the cell membrane and immature viruses bud off the cell. Core proteins are produced as part of long polypeptides, which must be cut into smaller fragments by the enzyme protease in order to form mature, functional proteins. Protease inhibitors bind to the site where protein cutting occurs, and so prevent the enzyme from releasing the individual core proteins. In this way the new viral particles are unable to mature or become infectious. (Adapted from ref. 108 with permission.)



pronounced restoration of immunity, attention has focused on the selective induction of augmented and effective HIV-specific immunity to supplement the imperfect and toxic antiretroviral treatments. Prophylactic immunization against viral disease was one of the major triumphs of twentieth-century medicine. The use of immunization to modulate chronic viral infection is still an unsubstantiated proposal.

Two approaches to augment HIV-specific immunity are under investigation: strategic treatment interruption (STI) and therapeutic vaccination. STI is based upon the hypothesis that after the arrest of progressive disease and the partial recovery of the immune system with potent antiretroviral therapy, the temporary interruption of therapy will release HIV antigen. This 'autoimmunization' followed by re-protection of the immune system with re-introduced chemotherapy can be performed in cycles until augmented immunity can change the natural history of infection.

The administration of very early chemotherapy during acute infection of patients or macaques with HIV or SIV, respectively, has been shown to change the natural history of disease, presumably by permitting the preservation and maturation of HIV-specific immunity without ongoing virus-mediated attack of CD4<sup>+</sup> cells<sup>103,104</sup>. Withdrawal of therapy after one or more cycles of treatment has resulted in well-controlled infection at least for periods of months.

But the results of STI have not been as encouraging in patients with chronic HIV infection. Here, autoimmunization of an impaired immune system is the objective but preservation of HIV-specific immunity is no longer an option. Moreover, in contrast to the acute situation, the immune response in chronic infection must contend with a complex mixture of genetic variants. This strategy remains under extensive investigation.

Therapeutic vaccination aims to deliver immunogens (substances capable of provoking an immune response) while the immune system is fully protected by potent antiretroviral chemotherapy. The precise responses that control HIV replication and the immunogens required to induce these have not been defined; nevertheless, quantitative assays for antigen-specific T-cell responses are proliferating and are being used to quantify correlates with degrees of protection<sup>105</sup>.

## Concluding perspective

In the past 15 years, HIV chemotherapy has progressed from preliminary trials with the first drug to a complex field of research and patient care that has markedly diminished mortality and morbidity in the developed world. This accomplishment defines several new challenges: (1) the discovery of drugs with increased potency, decreased toxicity and activity against drug-resistant virus; (2) the elimination of virus in poorly accessible tissue compartments and latent cellular reservoirs; (3) the induction of virus-specific immunity to supplement the benefits of chemotherapy; and (4) the identification of regimens that can benefit developing countries where the epidemic is most severe. Insights gained in drug discovery, pathogenesis and design of clinical trials for treatment of HIV will also be useful in addressing the serious and extensive viral epidemics of chronic hepatitis B and C. □

- Havilur, D. V. & Richman, D. D. Viral dynamics of HIV: implications for drug development and therapeutic strategies. *Ann. Intern. Med.* **124**, 984–994 (1996).
- Piatak, M. Jr *et al.* High levels of HIV-1 in plasma during all stages of infection determined by competitive PCR. *Science* **259**, 1749–1754 (1993).
- Wei, X. *et al.* Viral dynamics in human immunodeficiency virus type 1 infection. *Nature* **373**, 117–122 (1995).
- Ho, D. D. *et al.* Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature* **373**, 123–126 (1995).
- O'Brien, S. J., Nelson, G. W., Winkler, C. A. & Smith, M. W. Polygenic and multifactorial disease gene association in man: lessons from AIDS. *Ann. Rev. Genet.* **34**, 563–591 (2000).
- O'Brien, S. J. & Moore, J. P. The effect of genetic variation in chemokines and their receptors on HIV transmission and progression to AIDS. *Immunol. Rev.* **177**, 99–111 (2000).
- Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M. & Ho, D. D. HIV-1 dynamics in vivo: virion clearance rate, infected cell lifetime, and viral generation time. *Science* **271**, 1582–1586 (1996).
- Haase, A. T. *et al.* Quantitative image analysis of HIV-1 infection in lymphoid tissue. *Science* **274**, 985–989 (1996).

- Haase, A. T. Population biology of HIV-1 infection: viral and CD4<sup>+</sup> T cell demographics and dynamics in lymphatic tissues. *Ann. Rev. Immunol.* **17**, 625–656 (1999).
- Mellors, J. W. *et al.* Plasma viral load and CD4<sup>+</sup> lymphocytes as prognostic markers of HIV-1 infection. *Ann. Intern. Med.* **126**, 946–954 (1997).
- Carpenter, C. C. J. *et al.* Antiretroviral therapy in adults: updated recommendations of the International AIDS Society-USA panel. *J. Am. Med. Assoc.* **283**, 381–390 (2000).
- Veazey, R. S. *et al.* Gastrointestinal tract as a major site of CD4<sup>+</sup> T cell depletion and viral replication in SIV infection. *Science* **280**, 427–431 (1998).
- Harouse, J. M., Gettie, A., Tan, R. C. H., Blanchard, J. & Cheng-Mayer, C. Distinct pathogenic sequelae in rhesus macaques infected with CCR5 or CXCR4 utilizing SHIVs. *Science* **284**, 816–819 (1999).
- Haggerty, S. & Stevenson, M. Predominance of distinct viral genotypes in brain and lymph node compartments of HIV-1-infected individuals. *Viral Immunol.* **4**, 123–131 (1991).
- Epstein, L. G. *et al.* HIV-1 V3 domain variation in brain and spleen of children with AIDS: tissue-specific evolution within host-determined quasispecies. *Virology* **180**, 583–590 (1991).
- Wong, J. K. *et al.* In vivo compartmentalization of HIV: evidence from the examination of *pol* sequences from autopsy tissues. *J. Virol.* **70**, 2059–2071 (1997).
- Zhu, T. *et al.* Genetic characterization of human immunodeficiency virus type 1 in blood and genital secretions: evidence for viral compartmentalization and selection during sexual transmission. *J. Virol.* **70**, 3098–3107 (1996).
- Overbaugh, J., Anderson, R. J., Ndinya-Achola, J. O. & Kreiss, J. K. Distinct but related human immunodeficiency virus type 1 variant populations in genital secretions and blood. *AIDS Res. Hum. Retroviruses* **12**, 107–115 (1996).
- Wong, J. K. *et al.* Recovery of replication-competent HIV despite prolonged suppression of plasma viremia. *Science* **278**, 1291–1294 (1997).
- Finzi, D. *et al.* Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science* **278**, 1295–1300 (1997).
- Chun, T.-W. *et al.* Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy. *Proc. Natl Acad. Sci. USA* **94**, 13193–13197 (1997).
- Finzi, D. *et al.* Latent infection of CD4<sup>+</sup> T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. *Nature Med.* **5**, 512–517 (1999).
- Ramratnam, B. *et al.* The decay of the latent reservoir of replication-competent HIV-1 is inversely correlated with the extent of residual viral replication during prolonged anti-retroviral therapy. *Nature Med.* **6**, 82–85 (2000).
- Perelson, A. S. *et al.* Decay characteristics of HIV-1-infected compartments during combination therapy. *Nature* **387**, 188–191 (1997).
- Cavert, W. *et al.* Kinetics of response in lymphoid tissues to antiretroviral therapy of HIV-1 infection. *Science* **276**, 960–964 (1997).
- Zhang, Z.-Q. *et al.* Sexual transmission and propagation of SIV and HIV in resting and activated CD4<sup>+</sup> T cells. *Science* **286**, 1353–1357 (1999).
- Dornadula, G. *et al.* Residual HIV-1 RNA in blood plasma of patients taking suppressive highly active antiretroviral therapy. *J. Am. Med. Assoc.* **282**, 1627–1632 (1999).
- Günthard, H. F. *et al.* Evolution of envelope sequences of human immunodeficiency virus type 1 in cellular reservoirs in the setting of potent antiviral therapy. *J. Virol.* **73**, 9404–9412 (1999).
- Furtado, M. R. *et al.* Persistence of HIV-1 transcription in peripheral-blood mononuclear cells in patients receiving potent antiretroviral therapy. *N. Engl. J. Med.* **340**, 1614–1622 (1999).
- Sharkey, M. E. *et al.* Persistence of episomal HIV-1 infection intermediates in patients on highly active anti-retroviral therapy. *Nature Med.* **6**, 76–81 (2000).
- Yerly, S. *et al.* Time of initiation of antiretroviral therapy: impact on HIV-1 viraemia. *AIDS* **14**, 243–249 (2000).
- Fischer, M. *et al.* Residual HIV-RNA levels persist for up to 2.5 years in peripheral blood mononuclear cells of patients on potent antiretroviral therapy. *AIDS Res. Hum. Retroviruses* **16**, 1135–1140 (2000).
- Gulick, R. M. *et al.* Treatment with indinavir, zidovudine, and lamivudine in adults with human immunodeficiency virus infection and prior antiretroviral therapy. *N. Engl. J. Med.* **337**, 734–739 (1997).
- Autran, B. *et al.* Positive effects of combined antiretroviral therapy on CD4<sup>+</sup> T cell homeostasis and function in advanced HIV disease. *Science* **277**, 112–116 (1997).
- Pakker, N. G. *et al.* Biphasic kinetics of peripheral blood T cells after triple combination therapy in HIV-1 infection: a composite of redistribution and proliferation. *Nature Med.* **4**, 208–214 (1998).
- Lederman, M. M. *et al.* Immunologic responses associated with 12 weeks of combination antiretroviral therapy consisting of zidovudine, lamivudine, and zalcitabine: results of AIDS clinical trials group protocol 315. *J. Infect. Dis.* **178**, 70–79 (1998).
- Zhang, Z.-Q. *et al.* Kinetics of CD4<sup>+</sup> T cell repopulation of lymphoid tissues after treatment of HIV-1 infection. *Proc. Natl Acad. Sci. USA* **95**, 1154–1159 (1998).
- Bucy, R. P. *et al.* Initial increase in blood CD4<sup>+</sup> lymphocytes after HIV antiretroviral therapy reflects redistribution from lymphoid tissues. *J. Clin. Invest.* **103**, 1391–1398 (1999).
- McCune, J. M. *et al.* High prevalence of thymic tissue in adults with human immunodeficiency virus-1 infection. *J. Clin. Invest.* **101**, 2301–2308 (1998).
- Douek, D. C. *et al.* Changes in thymic function with age and during the treatment of HIV infection. *Nature* **396**, 690–695 (1999).
- Li, T. S. *et al.* Long-lasting recovery in CD4 T-cell function and viral-load reduction after highly active antiretroviral therapy in advanced HIV-1 disease. *Lancet* **351**, 1682–1686 (1998).
- Komanduri, K. V. *et al.* Restoration of cytomegalovirus-specific CD4<sup>+</sup> T-lymphocyte responses after ganciclovir and highly active antiretroviral therapy in individuals infected with HIV-1. *Nature Med.* **4**, 953–956 (1998).
- Havilur, D. V. *et al.* Effect of potent antiretroviral therapy on immune responses to *Mycobacterium avium* in human immunodeficiency virus-infected subjects. *J. Infect. Dis.* **182**, 1658–1663 (2000).
- Whitcup, S. M. Cytomegalovirus retinitis in the era of highly active antiretroviral therapy. *J. Am. Med. Assoc.* **283**, 653–657 (2000).
- Karavellas, M. P. *et al.* Incidence of immune recovery vitritis in cytomegalovirus retinitis patients following institution of successful highly active antiretroviral therapy. *J. Infect. Dis.* **179**, 697–700 (1999).
- Race, E. M. *et al.* Focal mycobacterial lymphadenitis following initiation of protease-inhibitor therapy in patients with advanced HIV-1 disease. *Lancet* **351**, 252–255 (1998).
- Kovacs, J. A. & Masur, H. Prophylaxis against opportunistic infections in patients with human immunodeficiency virus infection. *N. Engl. J. Med.* **342**, 1416–1429 (2000).



48. Mouton, Y. *et al.* Impact of protease inhibitors on AIDS-defining events and hospitalization in 10 French AIDS reference centres. *AIDS* **11**, F101–F105 (1997).
49. Centers for Disease Control and Prevention. HIV/AIDS Surveillance Rep. 10, 1–43 (1998).
50. Mocroft, A. *et al.* Changing patterns of mortality across Europe in patients infected with HIV-1. EuroSIDA Study Group. *Lancet* **352**, 1725–1730 (1998).
51. Palella, F. J. *et al.* Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. *N. Engl. J. Med.* **338**, 853–860 (1998).
52. Kohlstaedt, L. A., Wang, J., Friedman, J. M., Rice, P. A. & Steitz, T. A. Crystal structure at 3.5 Å resolution of HIV-1 reverse transcriptase complexed with an inhibitor. *Science* **256**, 1783–1790 (1992).
53. Rodgers, D. W. *et al.* The structure of unliganded reverse transcriptase from the human immunodeficiency virus type 1. *Proc. Natl Acad. Sci. USA* **92**, 1222–1226 (1995).
54. Haubrich, R. H. *et al.* The value of patient-reported adherence to antiretroviral therapy in predicting virologic and immunologic response. California Collaborative Treatment Group. *AIDS* **13**, 1099–1107 (1999).
55. Bangsberg, D. R. *et al.* Adherence to protease inhibitors, HIV-1 viral load, and development of drug resistance in an indigent population. *AIDS* **14**, 357–366 (2000).
56. Durant, J. *et al.* Importance of protease inhibitor plasma levels in HIV-infected patients treated with genotypic-guided therapy: pharmacological data from the Viradap Study. *AIDS* **14**, 1333–1339 (2000).
57. Gatti, G. *et al.* The relationship between ritonavir plasma levels and side-effects: implications for therapeutic drug monitoring. *AIDS* **13**, 2083–2089 (1999).
58. Flexner, C. HIV-protease inhibitors. *N. Engl. J. Med.* **338**, 1281–1292 (1998).
59. Enting, R. H. *et al.* Antiretroviral drugs and the central nervous system. *AIDS* **12**, 1941–1955 (1998).
60. Kim, R. B. *et al.* The drug transporter P-glycoprotein limits oral absorption and brain entry of HIV-1 protease inhibitors. *J. Clin. Invest.* **101**, 289–294 (1998).
61. Carr, A. *et al.* A syndrome of peripheral lipodystrophy, hyperlipidaemia and insulin resistance in patients receiving HIV protease inhibitors. *AIDS* **12**, F51–F58 (1998).
62. Safrin, S. & Grunfeld, C. Fat distribution and metabolic changes in patients with HIV infection. *AIDS* **13**, 2493–2505 (1999).
63. Carr, A., Miller, J., Law, M. & Cooper, D. A. A syndrome of lipodystrophy, lactic acidemia and liver dysfunction associated with HIV nucleoside analogue therapy: contribution to protease inhibitor-related lipodystrophy syndrome. *AIDS* **14**, F25–F32 (2000).
64. Mansky, L. M. & Temin, H. M. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *J. Virol.* **69**, 5087–5094 (1995).
65. Little, S. J. Transmission and prevalence of HIV resistance among treatment-naïve subjects. *Antivir. Ther.* **5**, 33–40 (2000).
66. Salomon, H. *et al.* Prevalence of HIV-1 resistant to antiretroviral drugs in 81 individuals newly infected by sexual contact or injecting drug use. Investigators of the Quebec Primary Infection Study. *AIDS* **14**, F17–F23 (2000).
67. Hirsch, M. S. *et al.* Antiretroviral drug resistance testing in adults with HIV infection. *J. Am. Med. Assoc.* **279**, 1984–1991 (1998).
68. Hirsch, M. S. *et al.* Antiretroviral drug resistance testing in adult HIV-1 infection. *J. Am. Med. Assoc.* **283**, 2417–2426 (2000).
69. Petropoulos, C. J. *et al.* A novel phenotypic drug susceptibility assay for human immunodeficiency virus type 1. *Antimicrob. Agents Chemother.* **44**, 920–928 (2000).
70. Vingerhoets, J. *et al.* The accuracy and reproducibility of high throughput genotypic and phenotypic HIV-1 resistance testing under EN45001 and CLIA accreditation labels. *Antivir. Ther.* **4**(Suppl. 1), 53 (1999).
71. DeGruttola, V. *et al.* The relation between baseline HIV drug resistance and response to antiretroviral therapy: re-analysis of retrospective and prospective studies using a standardized data analysis plan. *Antivir. Ther.* **5**, 41–48 (2000).
72. Clevenbergh, P. *et al.* Persisting long-term benefit of genotype-guided treatment for HIV- infected patients failing HAART. The Viradap Study: week 48 follow-up. *Antivir Ther* **5**, 65–70 (2000).
73. Baxter, J. D. *et al.* A randomized study of antiretroviral management based on plasma genotypic antiretroviral resistance testing in patients failing therapy. CPCRA 046 Study Team for the Terry Beinr Community Programs for Clinical Research on AIDS. *AIDS* **14**, F83–F93 (2000).
74. Condra, J. H. *et al.* Drug resistance and predicted virologic responses to human immunodeficiency virus type 1 protease inhibitor therapy. *J. Infect. Dis.* **182**, 758–765 (2000).
75. Kempf, D. *et al.* Identification of clinically relevant phenotypic and genotypic breakpoints for ABT-378/r in multiple PI-experience, NNRTI-naïve patients. *Antivir. Ther.* **5**(Suppl. 3), 70 (2000).
76. Andries, K. *et al.* R165335-TMC125, a novel non nucleoside reverse transcriptase inhibitor (NNRTI) with nanomolar activity against NNRTI resistant HIV strains. *AIDS* **14**(Suppl. 4), PL4.5 (2000).
77. Rice, W. G. *et al.* Inhibitors of HIV nucleocapsid protein zinc fingers as candidates for the treatment of AIDS. *Science* **270**, 1194–1197 (1995).
78. Hazuda, D. J. *et al.* Inhibitors of strand transfer that prevent integration and inhibit HIV-1 replication in cells. *Science* **287**, 646–650 (2000).
79. Moore, J. P. & Stevenson, M. New targets for inhibitors of HIV-1 replication. *Nature Rev. Mol. Cell Biol.* **1**, 40–49 (2000).
80. Chan, D. C., Fass, D., Berger, J. M. & Kim, P. S. Core structure of gp41 from the HIV envelope glycoprotein. *Cell* **89**, 263–273 (1997).
81. Kilby, J. M. *et al.* Potent suppression of HIV-1 replication in humans by T-20, a peptide inhibitor of gp41-mediated virus entry. *Nature Med.* **4**, 1302–1307 (1998).
82. Root, M. J., Kay, M. S. & Kim, P. S. Protein design of an HIV-1 entry inhibitor. *Science* **291**, 884–888 (2001).
83. Donzella, G. A. *et al.* AMD3100, a small molecule inhibitor of HIV-1 entry via the CXCR4 co-receptor. *Nature Med.* **4**, 72–77 (1999).
84. Baba, M. *et al.* A small-molecule, nonpeptide CCR5 antagonist with highly potent and selective anti-HIV-1 activity. *Proc. Natl Acad. Sci. USA* **96**, 5698–5703 (1999).
85. Bozzette, S. A. *et al.* The care of HIV-infected adults in the United States. *N. Engl. J. Med.* **339**, 1897–1904 (1998).
86. Hellinger, F. J. & Fleishman, J. A. Estimating the national cost of treating people with HIV disease: patient, payer, and provider data. *J. Acquir. Immune Defic. Syndr.* **24**, 182–188 (2000).
87. Gebo, K. A., Chaisson, R. E., Folkemer, J. G. & Moore, R. D. Costs of HIV medical care in the era of highly active antiretroviral therapy. *AIDS* **13**, 963–969 (1999).
88. Keiser, P. *et al.* Protease inhibitor-based therapy is associated with decreased HIV-related health care costs in men treated at a Veterans Administration hospital. *J. Acquir. Immune Defic. Syndr. Hum. Retrovirol.* **20**, 28–33 (1999).
89. Moore, R. D. Cost effectiveness of combination HIV therapy: 3 years later. *Pharmacoeconomics* **17**, 325–330 (2000).
90. Forsythe, S. S. The affordability of antiretroviral therapy in developing countries: what policymakers need to know. *AIDS* **12**(Suppl 2), S11–S18 (1998).
91. Shapiro, M. F. *et al.* Variations in the care of HIV-infected adults in the United States. *J. Am. Med. Assoc.* **281**, 2305–2315 (2000).
92. Marseille, E. *et al.* Cost effectiveness of single-dose nevirapine regimen for mothers and babies to decrease vertical HIV-1 transmission in sub-Saharan Africa. *Lancet* **354**, 803–809 (1999).
93. De Cock, K. M. *et al.* Prevention of mother-to-child HIV transmission in resource-poor countries: translating research into policy and practice. *J. Am. Med. Assoc.* **283**, 1175–1182 (2000).
94. Carmichael, A., Jin, X., Sissons, P. & Borysiewicz, L. Quantitative analysis of the human immunodeficiency virus type 1 (HIV-1)-specific cytotoxic T lymphocyte (CTL) response at different stages of HIV-1 infection: differential CTL responses to HIV-1 and Epstein-Barr virus in late disease. *J. Exp. Med.* **177**, 249–256 (1993).
95. Pantaleo, G. *et al.* Studies in subjects with long-term nonprogressive human immunodeficiency virus infection. *N. Engl. J. Med.* **332**, 209–216 (1995).
96. Rosenberg, E. S. *et al.* Vigorous HIV-1-specific CD4<sup>+</sup> T cell responses associated with control of viremia. *Science* **278**, 1447–1450 (1997).
97. Klein, M. R. *et al.* Kinetics of Gag-specific cytotoxic T lymphocyte responses during the clinical course of HIV-1 infection: a longitudinal analysis of rapid progressors and long-term asymptomatics. *J. Exp. Med.* **181**, 1365–1372 (1995).
98. Ogg, G. S. *et al.* Quantitation of HIV-1 specific cytotoxic T lymphocytes and plasma load of viral RNA. *Science* **279**, 2103–2106 (1998).
99. Koup, R. A. *et al.* Temporal association of cellular immune responses with the initial control of viremia in primary human immunodeficiency virus type 1 syndrome. *J. Virol.* **68**, 4650–4655 (1994).
100. Matano, T. *et al.* Administration of an anti-CD8 monoclonal antibody interferes with the clearance of chimeric simian/human immunodeficiency virus during primary infections of rhesus macaques. *J. Virol.* **72**, 164–169 (1998).
101. Schmitz, J. E. *et al.* Control of viremia in simian immunodeficiency virus infection by CD8<sup>+</sup> lymphocytes. *Science* **283**, 857–860 (1999).
102. Jin, X. *et al.* Dramatic rise in plasma viremia after CD8<sup>+</sup> T cell depletion in simian immunodeficiency virus-infected macaques. *J. Exp. Med.* **189**, 991–998 (1999).
103. Rosenberg, E. S. *et al.* Immune control of HIV-1 after early treatment of acute infection. *Nature* **407**, 523–526 (2000).
104. Lori, F. *et al.* Control of SIV rebound through structured treatment interruptions during early infection. *Science* **290**, 1591–1593 (2000).
105. Barouch, D. H. *et al.* Control of viremia and prevention of clinical AIDS in rhesus monkeys by cytokine-augmented DNA vaccination. *Science* **290**, 486–492 (2000).
106. Gulick, R. M. *et al.* 3-Year suppression of HIV viremia with indinavir, zidovudine and lamivudine. *Ann. Intern. Med.* **133**, 35–39 (2000).
107. Centers for Disease Control and Prevention. AIDS Surveillance—Trends <<http://www.cdc.gov/hiv/graphics/trends.htm>> Divisions of HIV/AIDS Prevention, National Center for HIV, STD and TB Prevention (11 December 2000).
108. F. Hoffmann-La Roche, Ltd. The multimedia lifecycle of HIV <<http://roche-hiv.com>> (1999).

#### Acknowledgements

A limited review of such an extensive field can achieve neither completeness nor balance. I apologize to my many colleagues whose contributions have not been explicitly acknowledged. D.R. is supported by grants from the UCSD Center for AIDS Research, the National Institutes of Health, and the Research Center for AIDS and HIV Infection of the Veterans Affairs San Diego Healthcare System.

# Challenges and opportunities for development of an AIDS vaccine

Gary J. Nabel

Vaccine Research Center, National Institutes of Allergy and Infectious Diseases, National Institutes of Health, 40 Convent Drive, Bethesda, Maryland 20892-3005, USA (e-mail: gnabel@nih.gov)

**Among the devastating consequences of AIDS has been its epidemic spread in the developing world. The disease has caused unprecedented suffering, debilitation, loss of life and disruption of family, social and economic stability. Because of the considerable expense and logistical difficulty in providing antiviral drugs to populations infected with the human immunodeficiency virus throughout the world, the biomedical community is looking towards vaccines to help solve this compelling problem.**

**T**he search for an AIDS vaccine began more than 15 years ago with great optimism and high expectations. With the identification of the human immunodeficiency virus (HIV) as the cause of AIDS, it seemed that a vaccine would follow closely behind. But despite a large concerted effort, the problem has proven more difficult than anticipated, and progress has not matched the initial hopes. Here I review the principal scientific obstacles confronting the development of an effective HIV vaccine, and I consider potential strategies to overcome these obstacles.

## The challenge

It is instructive to consider the circumstances that have contributed to past successes in vaccine development. The smallpox vaccine is among the most successful interventions in the history of medicine. Why, 200 years ago, without the benefit of modern biotechnology, did the smallpox vaccine succeed so readily while an AIDS vaccine remains elusive? The answer lies in an experiment of nature that provided, to an astute observer, a clear direction for smallpox vaccine development. In this classic story of scientific discovery, Edward Jenner noticed that milkmaids who had previously contracted cowpox were resistant to smallpox infection. This observation was the critical event leading to the finding that the cowpox virus cross-reacted immunologically with the smallpox virus and could therefore be used to protect against smallpox.

Jenner's milkmaids represented a protected population that provided the key information needed to develop the vaccine. Unfortunately, there are no significant large populations with well-defined resistance to HIV infection, and thus no immune parameters have been identified that correlate with protection. A cohort of exposed seronegative sex workers has been identified in Nairobi<sup>1</sup>, and it was hoped that this group could provide information about immune resistance to the virus. Unfortunately, protection is not always long lasting, the mechanism of their resistance has not been clearly identified, and thus they have not yielded definitive markers to guide vaccine development.

Development of an HIV vaccine is possible without the benefit of a correlate of immunity or a surrogate marker of protection in advance, but the road to its discovery will be considerably more arduous. The lack of immune correlates remains one of the more compelling challenges for the development of an AIDS vaccine (Table 1). The identification of immunogens that induce broad and long-lasting

immunity has been another critical hurdle. And the genetic diversity of the virus poses still another difficulty: HIV displays an unusual degree of diversity that confounds efforts to create a vaccine that is universally effective against the various clades and viral strains<sup>2</sup>. In contrast, the strain variation of other viruses, such as poliovirus, for which vaccines have succeeded, is relatively limited. For example, a vaccine from three strains of poliovirus has been used successfully worldwide<sup>3,4</sup>.

Progress in HIV vaccine research is also subject to the limitations of animal models. Lentiviral infection of non-human primates provides an unusually good animal model for viral vaccine development that has provided important insights into HIV immunopathogenesis. But despite many similarities in the symptoms and pathology, there remain distinctions that could affect vaccine efficacy. For example, the simian immunodeficiency viruses (SIV) differ significantly from their human counterparts. With regard to genomic organization, the SIV viral accessory protein Vpx is not found in HIV-1, the Vpu gene product of HIV-1 is not found in SIV (reviewed in ref. 5), and the functions of Vpr also differ between the two viruses. Although some HIV strains have been shown to cause disease in chimpanzees, these primates seem to be largely resistant to the CCR5-tropic primary isolates responsible for infection of most humans<sup>6</sup>, suggesting species differences in the host response to virus. The characteristics of molecular clones and laboratory-adapted viruses also differ from naturally infectious virus. Thus, despite the attractive features of the model, it is evident that human clinical studies will be needed for the development of effective vaccines. Such trials require the production of clinical-grade vaccines and requisite safety and toxicity studies, which pose additional challenges. As trials progress, the most important task is to choose the promising candidates for phase III clinical trials that test the efficacy of a vaccine based on limited information.

## The immune response to HIV vaccines

The immune system responds to infectious agents through the elaboration of a humoral response in which B lymphocytes produce secreted antibodies that interact with a variety of infectious agents, including bacteria, viruses, fungi and parasites. In addition, cell-mediated immune responses induce helper T lymphocytes that stimulate antibody responses and cytotoxic T lymphocytes (CTLs), which recognize processed antigens and lyse infected cells. Both of

these immune mechanisms can be manipulated in vaccine development and each has its advantages and disadvantages (Table 2).

#### Cell-mediated immunity

CTLs are also known as killer T cells because they recognize, bind and kill cells that display foreign antigens. Their role in protective immunity against viral infections has been well documented<sup>7,8</sup>. An immunogen that elicits a CTL response enables the recognition and elimination of infected cells, and a CTL response is highly desirable in an AIDS vaccine because CTLs can eliminate or reduce virus production by killing viral producer cells. The importance of CTL responses in limiting viral infection has been supported by human studies and by studies of non-human primates<sup>1,9,10</sup>. Potent CTL responses and resistance to HIV was seen in the exposed seronegative Nairobi sex workers described above<sup>1</sup>. And in experiments with chronic SIV-infected rhesus macaques, a marked increase in viral load was seen when CD8<sup>+</sup> T cells were depleted experimentally in these animals<sup>9,10</sup>.

Despite the importance of CTLs in containing viral replication, uncertainties remain as to whether cell-mediated immunity will be necessary and sufficient for a highly effective vaccine. Whether CTLs can be maintained in an active state for long periods of time or can expand rapidly enough to respond when an acute viral insult occurs is uncertain. Furthermore, vaccines that depend strictly on a CTL response may be countered by viral adaptations that allow the virus to evade detection by T cells. HIV, for example, has evolved mechanisms to disrupt the major histocompatibility complex (MHC) proteins, the antigen-presenting proteins that T cells rely on to recognize foreign antigens on the cell surface. One of these mechanisms involves the HIV gene product Nef, which downregulates expression of CD4 and class I MHC, cell-surface proteins that are essential to CTL recognition of viral antigens on viral producer cells<sup>11–13</sup>.

#### Humoral immunity

Because CTL responses alone are unlikely to provide complete protection, it will be important for an HIV vaccine to elicit neutralizing antibodies to the virus. Such antibodies are dependent on memory B cells, a long-lived cell population that can divide and differentiate into the antibody-producing plasma cell upon re-exposure to antigen, thus conferring long-term protection. Another advantage of antibodies is that they have the potential to inactivate virus before it has a chance to infect the cells of the host. Antibodies may also mobilize the inflammatory system, including the complement system, neutrophils and monocytes. Thus, even when an antibody does not directly neutralize the virus, there is potential by other antibody effector functions for amplification through the inflammatory system.

Although the antibody response should be a critical factor in antiviral immunity, one of the main hurdles for a highly effective HIV vaccine has been the development of immunogens that elicit broadly neutralizing antibodies. It has been possible to generate antibodies against the envelope protein of HIV, but such antibodies have limited efficacy. This problem arises, in part, because they neutralize laboratory-adapted strains but are not effective against primary isolates and are often strain-specific<sup>14</sup>. Although several broadly neutralizing monoclonal antibodies that neutralize the infectivity of different strains have been identified, it has not been possible to elicit this response with well-defined immunogens (reviewed in refs 15–18). The extensive genetic diversity among different strains and clades of HIV has created considerable difficulty in this regard. In addition, there have been suggestions that some antibodies may enhance infectivity<sup>19–21</sup>, and there are indeed precedents for this complication in vaccine development. In efforts to develop vaccines against the respiratory syncytial virus, some early vaccine candidates were found to exacerbate infection<sup>22,23</sup>. It is clear that antibody responses must be assessed carefully.

The ability of antibodies to confer protection at mucosal sites is also of great importance to the development of an effective AIDS vaccine. Several studies have investigated the ability of antibodies to

protect against infection from intravenous challenge<sup>14</sup> and through mucosal surfaces, with encouraging results. Using chimaeric SIV/HIV viruses (SHIVs) in non-human primates, Mascola and colleagues infused neutralizing antibodies into rhesus macaques to protect against vaginally transmitted infection<sup>24</sup>, while Baba and co-workers tested the effects of such antibodies in an oral mucosal exposure after birth<sup>25</sup>. Passive transfer of antibodies conferred protection against disease in both studies. Although relatively high concentrations of antibody ( $\geq 100 \mu\text{g ml}^{-1}$ ) were used — levels much higher than would ordinarily be achieved by vaccination — these studies showed that an appropriate serum antibody response might reduce infection at mucosal surfaces. It is unlikely that vaccination could achieve such a robust antibody response, although vaccines could also generate cellular immunity that might reduce the requisite neutralizing antibody concentration to protect against infection. It is also hoped that synergy between antibodies directed against different neutralizing determinants might reduce the concentration required for effective neutralization.

#### Replication-defective viral vaccines

The best example of protection against infection by a lentivirus (the retroviral genus of the HIV species) involves the use of a live-attenuated virus. Desrosiers and colleagues showed that monkeys infected with SIV, attenuated by the deletion of the viral gene *nef*, were not infected upon subsequent challenges with wild-type or *nef*-deleted SIV<sup>26</sup>. However, despite the initial excitement generated by these promising results, additional observations indicated potential hazards of this approach. It was eventually found that significant pathology occurred in both infant and adult macaques after exposure to attenuated SIVs<sup>27–29</sup>. In addition, there have been several reports of patients infected with HIV containing naturally occurring mutations in *nef* or the regulatory region of the long terminal repeat. Initially these patients exhibited less aggressive disease progression, but over time they were found to have reduced CD4 counts and increased viral loads<sup>30–32</sup>. Although live-attenuated viruses may prove ultimately to be effective and safe, many concerns remain to be addressed before these viruses will be acceptable for use in human clinical trials.

#### Multi-component immunity for an effective vaccine

In the absence of known immune correlates of protection, it would be most prudent to develop a vaccine that stimulates multiple components of the immune system<sup>33</sup>. The logical conclusion from the extensive literature on immune protection in lentiviral infection is that a combination of long-lived memory T cells, both CD8<sup>+</sup> CTLs and CD4<sup>+</sup> memory helper T cells, will probably be needed for a highly effective AIDS vaccine. At the same time, a strategy to induce broadly neutralizing antibodies will be required for highly effective, long-lasting immunity. It is now also evident that the  $\beta$ -chemokine receptors are necessary for HIV infection (reviewed in refs 34, 35). In fact, one of the most compelling examples of natural immunity to HIV is found in individuals with mutations in this receptor<sup>36,37</sup>. Thus, immunogens that induce antibodies which disrupt  $\beta$ -chemokine-receptor binding may prove to be useful.

In summary, although the immune correlates of protection remain unknown, there is evidence that cell-mediated immunity controls viral replication. At the same time, evidence from other successful vaccine approaches has indicated that long-term B-cell memory, through the antibody response, is crucial in immune protection. The challenge is to develop a vaccine that can elicit a broadly reactive T-cell response that is long lasting, and to identify antigens that will elicit the 'correct' (broadly neutralizing) antibody response.

#### Recombinant vector vaccines and adjuvants

A variety of vaccine strategies that use inactive viruses or individual viral components have been investigated<sup>38</sup>. The original and still most common immunogens used in vaccines are live-attenuated or inactivated viruses. Viruses attenuated from a pathogenic agent have proven



**Table 1 Summary of the main hurdles in development of an AIDS vaccine**

|                                                                                                     |
|-----------------------------------------------------------------------------------------------------|
| 1. Identification of immunogens that induce broad and long-lasting CTL immunity                     |
| 2. Definition of structures and immunization strategies that elicit broadly neutralizing antibodies |
| 3. Definition of immune correlates of protection in human or animal models                          |
| 4. Strategies to address HIV clade and strain diversity                                             |
| 5. Expansion of human clinical trials:                                                              |
| Clinical-grade vaccine production                                                                   |
| Diversity and duration of immune response                                                           |
| Prioritization and analysis of candidates                                                           |

safe and effective in many widely used vaccines, including the Sabin poliovirus and chickenpox vaccine. In these cases, a non-pathogenic attenuated virus elicits an immune response that cross-reacts immunologically with the virulent virus. More recently, replication-defective viral vectors unrelated to the pathogen have been used to deliver selected viral immunogens that might induce protective immunity. Another new approach has been the use of genetically engineered plasmid DNA to direct the synthesis of an immunogen within the host cells (reviewed in refs 39, 40). Combinations of these approaches, such as DNA priming and viral vector boosting, seem to be especially promising in animal models of vaccine development.

### Nonviral vectors

DNA vaccines contain a gene encoding a viral immunogen under the regulation of a eukaryotic enhancer/promoter and polyadenylation signals that confer appropriate expression of the viral immunogen. The various elements of DNA vaccines can be readily manipulated to optimize the level and duration of expression and the potency of the immunogen. When injected into muscle, cells surrounding the injection site internalize the plasmid and transport the DNA to the nucleus where transcription, translation and appropriate post-translational modification occur. Compared to recombinant protein vaccines produced in bacteria or yeast, the proteins expressed from a DNA vaccine are more likely to assume a native conformation, and their localized production facilitates uptake by antigen-presenting cells. Thus, antibodies generated against these immunogens are more likely to recognize and provide protection against native non-denatured proteins of the pathogen. In addition, antigen synthesis within cells will lead to more effective class I MHC processing and presentation that should stimulate CTL responses.

DNA vaccines could alleviate some potential disadvantages of live-attenuated virus vaccines such as the possibility of pathogenic infection and side effects of chronic immune stimulation. The feasibility of genetic immunization has been shown in several experimental model systems<sup>41–43</sup>. Furthermore, in rodents it has proven effective in inducing immunity to a variety of infectious diseases including influenza virus<sup>44</sup>, malaria<sup>45–48</sup>, tuberculosis<sup>49</sup>, Ebola virus<sup>50</sup>, rabies<sup>51</sup>, lymphocytic choriomeningitis virus<sup>52,53</sup> and herpes simplex virus<sup>54</sup>. Studies of non-human primates and humans have indicated that this approach is particularly effective in generating CTL responses<sup>55</sup>. A recent study showed that rhesus monkeys receiving a DNA vaccine augmented with a recombinant interleukin-2 plasmid developed potent CTL responses and did not develop clinical disease after a pathogenic SHIV challenge<sup>56</sup>. This was particularly significant not only for the success with a DNA vaccine strategy but also because vaccination prevented the appearance of disease symptoms.

Despite promising results in rodent models, DNA vaccinations have proven less effective in primates. It is likely that this technical issue can be overcome with time and increased experience, and several approaches have been taken to improve these vaccines. For example, more potent gene-expression strategies that use stronger enhancers or gene amplification have been explored. Another approach has been to modify the transcriptional and translational efficiency of foreign DNA by using codon choices preferred in the host species<sup>57,58</sup>. The use of human codons in some gene-based vectors, particularly from viruses and other infectious agents, increases

the level of production significantly, most likely by modifying RNA regulatory structures that prevent export from the nucleus to the cytoplasm, thus preventing effective full-length translation and transcription. By overcoming these blocks and optimizing expression, more antigen is available to present to the immune system, which may prove helpful in eliciting more effective immunity to HIV.

### Viral vectors

From efforts in vaccine development and gene therapy, new viral vectors have been advanced in recent years that may prove useful for AIDS vaccines. These include replication-defective forms of poxvirus vectors, including canarypox, fowlpox and modified vaccinia Ankara<sup>59–61</sup>. Highly immunogenic viruses, such as adenovirus, may also prove useful, particularly when strains are identified that are not reactive with antibodies commonly found in humans. Other vectors have progressed through pre-clinical development, including Sindbis<sup>62,63</sup> and Venezuelan equine encephalitis virus alphaviruses<sup>64,65</sup>. Replication-defective vectors from these viruses synthesize high levels of recombinant protein and can target delivery to dendritic cells. In addition, vectors have been identified that may allow persistent transgene expression to stimulate continuous T-cell activation. Advances in lentiviral vector development have shown that they can be modified to address a number of safety concerns<sup>66–68</sup>. Adeno-associated virus has also proven effective in achieving long-term gene expression and is the subject of research for AIDS vaccines<sup>69,70</sup>.

### Adjuvants

Key to the development of any successful vaccine is the use of adjuvants that augment immune responses to specific antigens. Adjuvants have traditionally been defined as substances used in combination with a specific antigen to elicit a more potent immunity than when the antigen is used by itself. A variety of adjuvants have been tested in animal models and human studies, and the subject is the topic of intense interest on which several insightful texts and reviews have been written<sup>71–75</sup>. With respect to HIV vaccines, several adjuvants have been tested in phase I clinical trials. These include polymers such as oligolysine, lipopeptides and polylactide co-polymers<sup>76</sup>. Traditional adjuvants such as aluminium phosphate or aluminium hydroxide, which are precipitated with the antigen, QS-21, MF59, monophosphoryl lipid A, mineral oil, mannose mono-oleate or incomplete Freund's adjuvant, purified protein derivative, keyhole limpet haemocyanin and bupivacaine (for DNA vaccines) have been analysed, with varying degrees of efficacy (summarized in ref. 76). In addition, a number of innovative technologies have been adapted, including the use of cytokine proteins or cytokine DNA expression vectors, immunostimulatory DNA sequences and the formulation of new complexes designed to create microparticles that can facilitate uptake of antigen-presenting cells.

The mechanisms by which these adjuvants work are not fully understood, but suggestions include: (1) preferential stimulation of specific T-cell subsets; (2) targeting of antigen to antigen-presenting cells; (3) direction of antigen into the MHC class I or class II pathways; (4) antigen deposition with slow release; and (5) stabilization of epitopes. A number of factors can affect the efficacy of adjuvants. These variables are related to the mode of administration, formulation of the adjuvant, species-specific responses to the immunostimulatory effects of the adjuvant, and the immune status of the host. So far, there is no clear preferential adjuvant for use in HIV vaccines. This has already been the topic of considerable investigation in phase I studies by the Vaccine Trials Network supported by the US National Institutes of Health and promises to be a continued area of important research for the development of highly effective vaccines.

### Choice of immunogens

HIV encodes more than 12 gene products, any of which might serve as targets for immune recognition. The synthesis of these viral proteins is regulated by viral transactivators and includes proteins derived from messenger RNAs synthesized from highly spliced viral

RNA, made early after the course of infection, and those derived from unspliced viral RNA, produced late in the viral life cycle. Several of these viral proteins contribute to the structure of virus and are synthesized in high quantities, such as the matrix and capsid proteins. Others give rise to regulatory proteins that modulate viral gene expression and are synthesized in lower quantities. Because CTLs are more likely than antibodies to be effective against internal viral proteins, attention has focused on the use of Gag proteins as immunogens for the CTL response. For Env, which is found on the surface of the virus, even though CTL responses are likely to be beneficial, the accessibility of this protein on intact virions would make it an attractive target for neutralizing antibodies. These two products of late, unspliced viral RNAs are widely considered to be important constituents of a highly effective HIV vaccine.

Among the highly spliced, early viral RNA products, Nef is expressed at high levels, and its importance in viral replication has been demonstrated<sup>77</sup>, although it is uncertain whether it will be an effective target for vaccine. Because it is expressed early in the virus life cycle, immune responses to this protein may serve to limit the burst size of virus from individual cells and could contribute to protective immunity. Finally, the Tat and Rev regulatory proteins have been the focus of study because of their potent regulatory activity. The Rev protein is found primarily in the nucleus but is not expressed at high levels; because it is not highly immunogenic it is not considered an attractive vaccine candidate. In contrast, the Tat protein, although also found in the nucleus at low levels, has been described in extracellular tissues, and several investigators have suggested that it may exert biological effects relevant to the pathogenesis of HIV disease (reviewed in ref. 78). Several laboratories have explored the potential utility of Tat as a constituent of an AIDS vaccine. Although reports have indicated that modified forms of Tat proteins can induce immune responses that may reduce viral replication in models using non-human primates<sup>79,80</sup>, it is not yet clear whether these models are relevant to natural infection in humans<sup>81</sup>. The protective effects induced by immunization with modified Tat proteins have not been observed with gene-based Tat delivery, and these findings have raised questions about the ultimate efficacy of Tat as an immunogen. Whether this difference is due to alternative immune responses to Tat by protein versus genetic immunization, to specific features of the prime models of infection, or to other undefined variables remains unclear and will require further investigation.

### Immune evasion by HIV

The challenges involved in developing an HIV vaccine go beyond issues of optimizing expression. Indeed, simply generating antibodies is no challenge. Rather, the difficulty lies in the identification of immunogens that stimulate the production of broadly neutralizing antibodies. There are a number of reasons why this has been so difficult, and they involve the many adaptations that HIV has evolved to thwart the immune system.

Because almost every infected individual generates an antibody response, HIV is a virus that has been selected in the presence of antibodies. The viral envelope is the part of the virus most accessible to the immune system. Thus, it has evolved under this selective pressure to evade immune detection (reviewed in ref. 82). This is accomplished in several ways, such as glycosylation of envelope proteins, and masking of critical parts of proteins including CD4 and co-receptor binding sites. In addition, the envelope itself seems to be conformationally active<sup>82</sup>. Thus, some structures that need to be recognized at the time the virus engages its receptor are not exposed to the immune system at a time when they could be accessible to B cells or antibody. It is also possible that the envelope has developed decoy mechanisms. For example, the protein includes epitopes that may attract an immune response that diverts recognition from highly conserved regions of the envelope critical for receptor binding and entry. HIV can also escape detection by the cellular immune response through multiple mechanisms, for

example, by Nef-mediated reduction in class I MHC expression or through complex gene regulation that permits latent infection of cells.

### Structure–function relationships in immunogen design

Clearly, gaining the advantage over HIV requires a more thorough understanding of viral envelope structure. X-ray crystallographic studies<sup>83–86</sup> have provided valuable information on the structures of the envelope proteins, gp120 and gp41. The crystal structure of gp120, for example, reveals numerous mechanisms of immune evasion, including conformational change, steric occlusion, islands of variation and a carbohydrate cloak<sup>82</sup>. Analysis of gp120 in complex with CD4, the primary virus receptor, for example, reveals that the CD4 binding site is recessed and contains several cavities. Many residues critical for antibodies directed against this region are not accessible in the CD4-bound conformation, although the rim of the conserved ‘Phe43 cavity’ may be accessible for neutralization<sup>82</sup>. It is likely that the native envelope protein will need to be altered as an immunogen to effectively present epitopes that can elicit neutralizing antibody responses. Other potential molecular targets include structures that may be exposed only transiently during fusion and entry. Alternatively, it may be possible to elicit one antibody that induces a conformational change in the envelope spike that would expose otherwise cryptic sites (for example, normally masked chemokine-receptor binding surfaces on gp120) to attack by a second antibody.

The viral envelope has intrinsic immunogenicity, and when it is injected into its recipient, it tends to induce a certain stereotypical immune response. Different components of the virus generate different responses. For example, when mice are injected with DNA encoding gp160, a CTL response is readily generated. However, if antibody production is examined in the same mice, the titre is found to be very low, whereas another viral gene, *nef*, induces high titre antibodies and low CTL activity. It is important to understand the genetic and structural bases for divergent immune responses, for example, by the analysis of the immunogenicity of diverse mutant envelope proteins. The ultimate goal is to combine immunogenicity information with an analysis of the physical structure of these mutant proteins. Although this information may differ depending on the vector and/or adjuvant, such understanding of structure and its relation to immunogenicity may help to identify promising vaccine candidates.

Another structural approach to vaccine design has derived from analysis of the mechanism of viral fusion and the HIV gp41 region

**Table 2 Possible approaches to vaccine development**

|                                      | Advantages                                                                         | Disadvantages                                              |
|--------------------------------------|------------------------------------------------------------------------------------|------------------------------------------------------------|
| CTL-eliciting vaccines               | Recognition of virally infected cells                                              | Requirement for active, long-term memory cells             |
|                                      | Multiple linear epitopes                                                           | Inability to recognize virus in absence of MHC             |
|                                      | Elimination of virus production                                                    | Down-modulation of MHC by virus                            |
|                                      | Possible effect on latent reservoir                                                |                                                            |
| Antibody-eliciting vaccines          | Neutralization of virus                                                            | Inability to generate broadly neutralizing antibodies      |
|                                      | Ability to prevent new infection of cells                                          | Evolution of resistant strains                             |
|                                      | Activation of inflammatory response (complement system, neutrophils and monocytes) | Antibodies may enhance infectivity to exacerbate infection |
| Replication-defective viral vaccines | Persistent antigen expression                                                      | Safety and inadvertent infection                           |
|                                      | Induction of cell-mediated and humoral immunity                                    | Enhanced replication during immune suppression             |
|                                      | Interference                                                                       | Consequences of persistent immune stimulation              |
|                                      |                                                                                    | Increased rates of integration                             |

that contains structures relevant to this process. Fusion requires two triple helical coiled-coil regions that fold like a hairpin to generate a six helical bundle<sup>85,86</sup>. The helical coiled-coil structure facilitates insertion of the fusion peptide into its target cell. This structure is not unique to HIV, as it is found in a variety of viruses, including influenza virus, murine leukaemia viruses, SIV and respiratory syncytial virus, and within eukaryotic cells in the soluble NSF attachment receptor (SNARE) involved in vesicle fusion (summarized in ref. 87).

Matthews, Hunter and Bolognesi<sup>88</sup> and Kim and co-workers<sup>89</sup> have shown that a peptide that interferes with the helical coiled-coil structure of HIV gp41 disrupts viral fusion and decreases viral load in human clinical studies. Efforts have been made to focus on this structure as a target for neutralization by antibodies, but, unfortunately, antibodies against this structure do not seem to neutralize the virus. It is likely that once the six-helical coiled-coil structure has formed, it is not accessible to antibodies. Thus, another approach has been to target intermediate structures that form before the hairpin has been generated. Nunberg and co-workers used this approach in experiments where mice were immunized with complexes of Env-expressing cells admixed with CD4<sup>+</sup>/CCR5<sup>+</sup> target cells in an effort to form fusion intermediates<sup>90</sup>. Although promising in the reported study, this method has been difficult to apply to primate models and will prove challenging for large-scale vaccine production. Thus, many laboratories continue to address the critical problem of immunization with fusion intermediates using alternative molecular approaches. It is not yet known whether such envelope proteins can be identified, but through a combination of DNA technology and structural information, it will be possible to rapidly create and survey a variety of immunogens that are both informative and useful in vaccine development.

### Current and future directions for AIDS vaccine research

The challenges faced in the development of an AIDS vaccine are similar to those that remain for a number of medically important infectious diseases, including malaria and tuberculosis. In all cases, safe, effective, broad-spectrum vaccines are needed that generate long-term immune memory and protection at the sites of infection, particularly at mucosal sites where the virus gains entry into the body. Because the vaccines are needed in parts of the world where the most modern medical care is not necessarily available, it will also be necessary to have simple, transportable vectors that can be administered easily.

HIV vaccines have been evaluated so far in over 70 phase I (dose-escalation safety and toxicity), five phase II (expanded safety and dose optimization) and two phase III (efficacy) clinical trials. These studies have evaluated safety and immunogenicity of preventive vaccines in more than 3,500 subjects. Among them, several envelope proteins, which have been made in insect, yeast or mammalian cells with recombinant DNA technology and which encode gp120 or gp160, have been administered with different adjuvants. Peptides derived from the envelope V3 loop or Gag have been delivered as conjugates to oligolysine backbones, as lipopeptide conjugates, with alum, or as fusion proteins with the self-assembling yeast protein Ty as a particle. Alternative sites of delivery have been analysed, including intramuscular, oral and rectal routes. In addition, gene-based vectors, such as vaccinia, canarypox and salmonella, as well as nucleic acid-based vaccines, have been tested. These studies are reviewed in detail elsewhere<sup>76,91,92</sup>. No significant safety concerns have arisen. Although neutralizing antibody responses have been detected, their activity on primary CCR5-tropic isolates has been minimal and has largely been strain-specific<sup>93,94</sup>. CD8<sup>+</sup> CTL responses have also been found that are durable, with some cross-clade reactivity<sup>95</sup>, although the consistency of such responses is not yet optimal. Data on the efficacy of any of these vaccines in human populations are not yet available. Most researchers believe that a highly promising ideal vaccine candidate is not at hand, and that it will be necessary to evaluate additional candidates that generate strong, broad and long-lasting CTL and neutralizing antibody responses.

There are numerous scientific opportunities to enhance vaccine development. Methods must be developed to enhance the immunogenicity of specific HIV peptides and to enhance antibody responses to regions of HIV proteins that are naturally protected from immune detection. These regions are generally the most conserved, so antibodies recognizing them would be most likely to be more broadly effective across the various clades and strains of HIV. Improved understanding of adjuvants is also needed.

Another important tool for vaccine development is the rapidly evolving field of genomics. An understanding of the human genome as well as the viral genome will help to identify genetic factors that determine immunological responders and confer immunological resistance. Analysis of the gene-expression profiles and linkage studies with polymorphic human genetic markers during vaccine trials and in association with HIV infection will facilitate identification of genes that confer resistance to infection, and may help to determine which vaccines are best suited to specific populations.

Considerable work remains to be done in the area of translation from animal studies to clinical trials. New vectors and more efficient methods for the production of vectors are needed. The most important aspect of translation is to understand the response of the immune system to vaccines. The original antibody and CTL assays have been invaluable in this regard, but they are relatively time consuming and cumbersome. Newer and more efficient technologies are being developed that provide detailed information about T-cell-receptor specificities and cytokine profiles<sup>96</sup>. These will be essential for identification and implementation of new vaccine candidates.

Human clinical studies remain the critical link between laboratory research and an effective vaccine. Vital to this effort is an informative network of clinical trials, both in the developed and developing world. Also crucial is the successful collaboration of the public and private sector in establishing the knowledge base, production expertise and vaccine distribution network that will someday be required for a successful vaccine. The search for an HIV vaccine has been slow and at times frustrating, but the resolve of the biomedical research community to address this problem has grown. Although the solution is not yet at hand, progress is tangible and encouraging results now develop on a regular basis. This renewed commitment and these advances in the science of AIDS vaccine development will make it possible to meet this unprecedented challenge. □

- Rowland-Jones, S. L. *et al.* Cytotoxic T cell response to multiple conserved HIV epitopes in HIV-resistant prostitutes in Nairobi. *J. Clin. Invest.* **102**, 1758–1765 (1998).
- Saag, M. S. *et al.* Extensive variation of human immunodeficiency virus type-1 *in vivo*. *Nature* **334**, 440–444 (1988).
- Melnick, J. L. in *Fields Virology* (eds Fields, B. N., Knipe, D. M. & Howley, P. M.) 655–712 (Lippincott-Raven, Philadelphia, 1996).
- Chanock, R. M. Reminiscences of Albert Sabin and his successful strategy for the development of the live oral poliovirus vaccine. *Proc. Assoc. Am. Phys.* **108**, 117–126 (1996).
- Luciw, P. A. in *Fields Virology* (eds Fields, B. N., Knipe, D. M. & Howley, P. M.) 1881–1952 (Lippincott-Raven, Philadelphia, 1996).
- Novembre, F. J. *et al.* Development of AIDS in a chimpanzee infected with human immunodeficiency virus type 1. *J. Virol.* **71**, 4086–4091 (1997).
- Zinkernagel, R. & Doherty, P. Major transplantation antigens, viruses, and specificity of surveillance T cells. *Contemp. Top. Immunobiol.* **7**, 179–220 (1977).
- Zinkernagel, R. & Doherty, P. MHC-restricted cytotoxic T cells: studies on the biological role of polymorphic major transplantation antigens determining T-cell restriction-specificity, function, and responsiveness. *Adv. Immunol.* **27**, 51–177 (2000).
- Jin, X. *et al.* Dramatic rise in plasma viremia after CD8<sup>+</sup> T cell depletion in simian immunodeficiency virus-infected macaques. *J. Exp. Med.* **189**, 991–998 (1999).
- Schmitz, J. E. *et al.* Control of viremia in simian immunodeficiency virus infection by CD8<sup>+</sup> lymphocytes. *Science* **283**, 857–860 (1999).
- Collins, K. L., Chen, B. K., Kalams, S. A., Walker, B. D. & Baltimore, D. HIV-1 Nef protein protects infected primary cells against killing by cytotoxic T lymphocytes. *Nature* **391**, 397–401 (1998).
- Schwartz, O., Marechal, V., Le Gall, S., Lemonnier, F. & Heard, J. M. Endocytosis of major histocompatibility complex class I molecules is induced by the HIV-1 Nef protein. *Nature Med.* **2**, 338–342 (1996).
- Cullen, B. R. HIV-1 auxiliary proteins: making connections in a dying cell. *Cell* **93**, 685–692 (1998).
- Mascola, J. R. *et al.* Protection of macaques against pathogenic simian/human immunodeficiency virus 89.6PD by passive transfer of neutralizing antibodies. *J. Virol.* **73**, 4009–4018 (1999).
- Parren, P. W., Burton, D. R. & Sattentau, Q. J. HIV-1 antibody—debris or virion? *Nature Med.* **3**, 366–367 (1997).
- Sattentau, Q. J. Neutralization of HIV-1 by antibody. *Curr. Opin. Immunol.* **8**, 540–545 (1996).



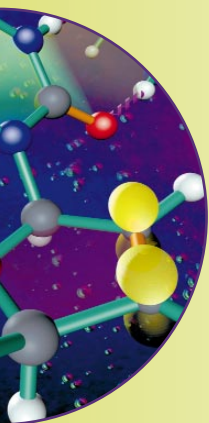
17. Burton, D. R. A vaccine for HIV type 1: the antibody perspective. *Proc. Natl Acad. Sci. USA* **94**, 10018–10023 (1997).
18. Burton, D. R. & Montefiori, D. C. The antibody response in HIV-1 infection. *AIDS* **11**(Suppl. A), S87–S98 (1997).
19. Morens, D. M. Antibody-dependent enhancement of infection and the pathogenesis of viral disease. *Clin. Infect. Dis.* **19**, 500–512 (1994).
20. Burke, D. S. Human HIV vaccine trials: does antibody-dependent enhancement pose a genuine risk? *Perspect. Biol. Med.* **35**, 511–530 (1992).
21. Mascola, J. R. *et al.* Summary report: workshop on the potential risks of antibody-dependent enhancement in human HIV vaccine trials. *AIDS Res. Hum. Retroviruses* **9**, 1175–1184 (1993).
22. Johnson, P. R. & Collins, P. L. Sequence comparison of the phosphoprotein mRNAs of antigenic subgroups A and B of human respiratory syncytial virus identifies a highly divergent domain in the predicted protein. *J. Gen. Virol.* **71**, 481–485 (1990).
23. Kim, H. W. *et al.* Respiratory syncytial virus disease in infants despite prior administration of antigenic inactivated vaccine. *Am. J. Epidemiol.* **89**, 422–434 (1969).
24. Mascola, J. R. *et al.* Protection of macaques against vaginal transmission of a pathogenic HIV-1/SIV chimeric virus by passive infusion of neutralizing antibodies. *Nature Med.* **6**, 207–210 (2000).
25. Baba, T. W. *et al.* Human neutralizing monoclonal antibodies of the IgG1 subtype protect against mucosal simian-human immunodeficiency virus infection. *Nature Med.* **6**, 200–206 (2000).
26. Daniel, M. D., Kirchhoff, F., Czajak, S. C., Sehgal, P. K. & Desrosiers, R. C. Protective effects of a live attenuated SIV vaccine with a deletion in the nef gene. *Science* **258**, 1938–1941 (1992).
27. Baba, T. W. *et al.* Pathogenicity of live, attenuated SIV after mucosal infection of neonatal macaques. *Science* **267**, 1820–1825 (1995).
28. Baba, T. W. *et al.* Live attenuated, multiply deleted simian immunodeficiency virus causes AIDS in infant and adult macaques. *Nature Med.* **5**, 194–203 (1999).
29. Wyand, M. S. *et al.* Protection by live, attenuated simian immunodeficiency virus against heterologous challenge. *J. Virol.* **73**, 8356–8363 (1999).
30. Deacon, N. J. *et al.* Genomic structure of an attenuated quasi species of HIV-1 from a blood transfusion donor and recipients. *Science* **270**, 988–991 (1995).
31. Greenough, T. C., Sullivan, J. L. & Desrosiers, R. C. Declining CD4 T-cell counts in a person infected with nef-deleted HIV-1. *N. Engl. J. Med.* **340**, 236–237 (1999).
32. Learmont, J. C. *et al.* Immunologic and virologic status after 14 to 18 years of infection with an attenuated strain of HIV-1. *N. Engl. J. Med.* **340**, 1715–1722 (1999).
33. Dittmer, U., Brooks, D. M. & Hasenkrug, K. J. Requirement for multiple lymphocyte subsets in protection by a live attenuated vaccine against retroviral infection. *Nature Med.* **5**, 189–193 (1999).
34. Gallo, R. C., Garzino-Demo, A. & DeVico, A. L. HIV infection and pathogenesis: what about chemokines? *J. Clin. Immunol.* **19**, 293–299 (1999).
35. Garzino-Demo, A., DeVico, A. L., Cocchi, F. & Gallo, R. C. Beta-chemokines and protection from HIV type 1 disease. *AIDS Res. Hum. Retroviruses* **14**(Suppl. 2), S177–S184 (1998).
36. Dean, M. *et al.* Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CCR5 structural gene. *Science* **273**, 1856–1862 (1996).
37. Samson, M. *et al.* Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature* **382**, 722–725 (1996).
38. Letvin, N. L. Progress in the development of an HIV-1 vaccine. *Science* **280**, 1875–1880 (1998).
39. Donnelly, J. J., Ulmer, J. B., Shiver, J. W. & Liu, M. A. DNA vaccines. *Annu. Rev. Immunol.* **15**, 617–648 (1997).
40. Tighe, H., Corr, M., Roman, M. & Raz, E. Gene vaccination: plasmid DNA is more than just a blueprint. *Immunol. Today* **19**, 89–97 (1998).
41. Tang, D. C., DeVit, M. & Johnston, S. A. Genetic immunization is a simple method for eliciting an immune response. *Nature* **356**, 152–154 (1992).
42. Liu, M. A. *et al.* Vaccination of mice and nonhuman primates using HIV-gene-containing DNA. *Antibiot. Chemother.* **48**, 100–104 (1996).
43. Wang, B. *et al.* DNA inoculation induces protective in vivo immune responses against cellular challenge with HIV-1 antigen-expressing cells. *AIDS Res. Hum. Retroviruses* **10**(Suppl. 2), S35–S41 (1994).
44. Ulmer, J. B. *et al.* Heterologous protection against influenza by injection of DNA encoding a viral protein. *Science* **259**, 1745–1749 (1993).
45. Becker, S. I. *et al.* Protection of mice against Plasmodium yoelii sporozoite challenge with P. yoelii merozoite surface protein 1 DNA vaccines. *Infect. Immun.* **66**, 3457–3461 (1998).
46. Doolan, D. L. *et al.* Circumventing genetic restriction of protection against malaria with multigene DNA immunization: CD8<sup>+</sup> cell-, interferon gamma-, and nitric oxide-dependent immunity. *J. Exp. Med.* **183**, 1739–1746 (1996).
47. Gardner, M. J. *et al.* DNA vaccines against malaria: immunogenicity and protection in a rodent model. *J. Pharm. Sci.* **85**, 1294–1300 (1996).
48. Sedegah, M., Hedstrom, R., Hobart, P. & Hoffman, S. L. Protection against malaria by immunization with plasmid DNA encoding circumsporozoite protein. *Proc. Natl Acad. Sci. USA* **91**, 9866–9870 (1994).
49. Tascon, R. C. *et al.* Vaccination against tuberculosis by injection. *Nature Med.* **2**, 888–892 (1996).
50. Xu, L. *et al.* Immunization for Ebola virus infection. *Nature Med.* **4**, 37–42 (1998).
51. Lodmell, D. L. *et al.* DNA immunization protects nonhuman primates against rabies virus. *Nature Med.* **4**, 949–952 (1998).
52. Martins, L. P., Lau, L. L., Asano, M. S. & Ahmed, R. DNA vaccination against persistent viral infection. *J. Virol.* **69**, 2574–2582 (1995).
53. Yokoyama, M., Zhang, J. & Whitton, J. L. DNA immunization confers protection against lethal lymphocytic choriomeningitis virus infection. *J. Virol.* **69**, 2684–2688 (1995).
54. Manickan, E., Yu, Z., Rouse, R. J., Wire, W. S. & Rouse, B. T. Induction of protective immunity against herpes simplex virus with DNA encoding the immediate early protein ICP 27. *Viral Immunol.* **8**, 53–61 (1995).
55. Wang, R. *et al.* Induction of antigen-specific cytotoxic T lymphocytes in humans by a malaria DNA vaccine. *Science* **282**, 476–480 (1998).
56. Barouch, D. H. *et al.* Control of viremia and prevention of clinical AIDS in rhesus monkeys by cytokine-augmented DNA vaccination. *Science* **290**, 486–492 (2000).
57. Schwartz, S. *et al.* Mutational inactivation of an inhibitory sequence in human immunodeficiency virus type 1 results in Rev-independent gag expression. *J. Virol.* **66**, 7176–7182 (1992).
58. Andre, S. *et al.* Increased immune response elicited by DNA vaccination with a synthetic gp120 sequence with optimized codon usage. *J. Virol.* **72**, 1497–1503 (1998).
59. Tartaglia, J. *et al.* Canarypox virus-based vaccines: prime-boost strategies to induce cell-mediated and humoral immunity against HIV. *AIDS Res. Hum. Retroviruses* **14**(Suppl. 3), S291–S298 (1998).
60. Moss, B. *et al.* Host range restricted, non-replicating vaccinia virus vectors as vaccine candidates. *Adv. Exp. Med. Biol.* **397**, 7–13 (1996).
61. Tartaglia, J., Pincus, S. & Paoletti, E. Poxvirus-based vectors as vaccine candidates. *Crit. Rev. Immunol.* **10**, 13–30 (1990).
62. Wahlfors, J. J., Zullo, S. A., Loimas, S., Nelson, D. M. & Morgan, R. A. Evaluation of recombinant alphaviruses as vectors in gene therapy. *Gene Ther.* **7**, 472–480 (2000).
63. Huang, H. V. Sindbis virus vectors for expression in animal cells. *Curr. Opin. Biotechnol.* **7**, 531–535 (1996).
64. Davis, N. L. *et al.* Vaccination of macaques against pathogenic simian immunodeficiency virus with Venezuelan equine encephalitis virus replicon particles. *J. Virol.* **74**, 371–378 (2000).
65. Caley, I. J. *et al.* Venezuelan equine encephalitis virus vectors expressing HIV-1 proteins: vector design strategies for improved vaccine efficacy. *Vaccine* **17**, 3124–3135 (1999).
66. Naldini, L. *et al.* In vivo gene delivery and stable transduction of nondividing cells by a lentiviral vector. *Science* **272**, 263–267 (1996).
67. Zufferey, R. *et al.* Self-inactivating lentivirus vector for safe and efficient in vivo gene delivery. *J. Virol.* **72**, 9873–9880 (1998).
68. Trono, D. Lentiviral vectors: turning a deadly foe into a therapeutic agent. *Gene Ther.* **7**, 20–23 (2000).
69. Monahan, P. E. & Samulski, R. J. AAV vectors: is clinical success on the horizon? *Gene Ther.* **7**, 24–30 (2000).
70. Liu, X. L., Clark, K. R. & Johnson, P. R. Production of recombinant adeno-associated virus vectors using a packaging cell line and a hybrid recombinant adenovirus. *Gene Ther.* **6**, 293–299 (1999).
71. Spriggs, D. R. & Koff, W. C. *Topics in Vaccine Adjuvant Research* (CRC, Boca Raton, FL, 1991).
72. Powell, M. F. & Newman, M. J. *Vaccine Design: The Subunit and Adjuvant Approach* (Plenum, New York, NY, 1995).
73. Cooper, P. D. in *Strategies in Vaccine Design* (ed. Ada, G. L.) 125–158 (Landes, Austin, TX, 1994).
74. Cox, J. C. & Coulter, A. R. Adjuvants—a classification and review of their modes of action. *Vaccine* **15**, 248–256 (1997).
75. O'Hagan, D. T. *Vaccine Adjuvants: Preparation Methods and Research Protocols* (Humana Press, Totowa, NJ, 2000).
76. Graham, B. in *HIV Molecular Immunology Database* (eds Korber, B. T. *et al.*) (Los Alamos National Laboratory: Theoretical Biology and Biophysics, Los Alamos, NM, 2000).
77. Kestler, H. W. III *et al.* Importance of the nef gene for maintenance of high virus loads and for development of AIDS. *Cell* **65**, 651–662 (1991).
78. Emerman, M. & Malim, M. H. HIV-1 regulatory/accessory genes: keys to unraveling viral and host cell biology. *Science* **280**, 1880–1884 (1998).
79. Cafaro, A. *et al.* Control of SHIV-89.6P-infection of cynomolgus monkeys by HIV-1 Tat protein vaccine. *Nature Med.* **5**, 643–650 (1999).
80. Pauza, C. D. *et al.* Vaccination with tat toxoid attenuates disease in simian/HIV-challenged macaques. *Proc. Natl Acad. Sci. USA* **97**, 3515–3519 (2000).
81. Travis, J. A controversial shot in the arm. *Science News* **156**, 300–303 (1999).
82. Wyatt, L. & Sodroski, J. The HIV-1 envelope glycoproteins: fusogens, antigens, and immunogens. *Science* **280**, 1884–1888 (1998).
83. Kwong, P. D. *et al.* Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature* **393**, 648–659 (1998).
84. Kwong, P. D. Structures of HIV-1 gp120 envelope glycoproteins from laboratory-adapted and primary isolates. *Structure* **8**, 1329–1339 (2000).
85. Chan, D. C., Fass, D., Berger, J. M. & Kim, P. S. Core structure of gp41 from the HIV envelope glycoprotein. *Cell* **89**, 263–273 (1997).
86. Weissenhorn, W., Dessen, A., Harrison, S. C., Skehel, J. J. & Wiley, D. C. Atomic structure of the ectodomain from HIV-1 gp41. *Nature* **387**, 426–430 (1997).
87. Weissenhorn, W., Carfi, A., Lee, K. H., Skehel, J. J. & Wiley, D. C. Crystal structure of the Ebola virus membrane fusion subunit, GP2, from the envelope glycoprotein ectodomain. *Mol. Cell* **2**, 605–616 (1998).
88. Kilby, J. M. *et al.* Potent suppression of HIV-1 replication in humans by T-20, a peptide inhibitor of gp41-mediated virus entry. *Nature Med.* **4**, 1302–1307 (1998).
89. Eckert, D. M., Malashkevich, V. N., Hong, L. H., Carr, P. A. & Kim, P. S. Inhibiting HIV-1 entry: discovery of D-peptide inhibitors that target the gp41 coiled-coil pocket. *Cell* **99**, 103–115 (1999).
90. LaCasse, R. A. *et al.* Fusion-competent vaccines: broad neutralization of primary isolates of HIV. *Science* **283**, 357–362 (1999).
91. Graham, B. S. & Karzon, D. T. in *Textbook of AIDS Medicine* (eds Merigan, T. C. Jr, Bartlett, J. G. & Bolognesi, D.) 689–724 (Lippincott, Williams, and Wilkins, Baltimore, MD, 1998).
92. Dolin, R. Human studies in the development of human immunodeficiency virus vaccines. *J. Infect. Dis.* **172**, 1175–1183 (1995).
93. Wrin, T., Loh, T. P., Vennari, J. C., Schuitemaker, H. & Nunberg, J. H. Adaptation to persistent growth in the H9 cell line renders a primary isolate of human immunodeficiency virus type 1 sensitive to neutralization by vaccine sera. *J. Virol.* **69**, 39–48 (1995).
94. Mascola, J. R. *et al.* Immunization with envelope subunit vaccine products elicits neutralizing antibodies against laboratory-adapted but not primary isolates of human immunodeficiency virus type 1. *J. Infect. Dis.* **173**, 340–348 (1996).
95. Ferrari, G. *et al.* Clade B-based HIV-1 vaccines elicit cross-clade cytotoxic T lymphocyte reactivities in uninfected volunteers. *Proc. Natl Acad. Sci. USA* **94**, 1396–1401 (1997).
96. Pitcher, C. J. *et al.* HIV-1-specific CD4<sup>+</sup> T cells are detectable in most individuals with active HIV-1 infection, but decline with prolonged viral suppression. *Nature Med.* **5**, 518–525 (1999).

#### Acknowledgements

I thank B. Graham, P. Kwong, J. Mascola and R. Wyatt for helpful discussions and comments and T. Suhana, A. Tislerics and N. Barrett for assistance with preparation of the manuscript.

# BRISTOL-MYERS SQUIBB AND HIV/AIDS:

## Basic Science, Clinical Development, Partnerships



Worldwide, more than 36 million people are estimated to be living with HIV/AIDS.<sup>1</sup> Through 2000, cumulative HIV/AIDS-associated deaths worldwide numbered approximately 21.8 million – 17.5 million adults and 4.3 million children under the age of 15.<sup>2</sup> These numbers and the massive human tragedy that they signify underscore the urgent need to discover and develop more effective antiretroviral therapies and make them available to patients as quickly and as broadly as possible.

Nearly twenty years have passed since acquired immunodeficiency syndrome was recognized and the human immunodeficiency virus was identified as the causative agent. Still, there is neither a cure nor a vaccine for HIV/AIDS. Important advances, however, have been made towards suppressing the virus, bolstering the immune system and extending and enhancing the lives of patients through combination antiretroviral therapy. Since the 1980s, researchers at Bristol-Myers Squibb have been a driving force in the development of antiretroviral treatments, having successfully ushered *Zerit*® (stavudine) and *Videx*® (didanosine), two nucleoside analogues, through clinical development and registration. Today, our research pipeline spans traditional and novel classes of antiretroviral drugs. While our clinical team focuses on the late stage development of the protease inhibitor BMS-232632, our discovery team is singularly charged with identifying novel classes of antiretrovirals.

Current antiviral drug combinations successfully suppress HIV in the majority of patients.<sup>3</sup> In the rest, the drugs fail to provide a sustained response. Researchers have long attributed drug failures to HIV's ability to develop drug resistance. Recent studies, however, have unearthed non-adherence to therapy regimens as the primary cause of treatment failure.<sup>4</sup> As our researchers continue to closely study resistance patterns, they also continue to seek to make our therapies more convenient and tolerable for patients. For example, just this past year, an improved formulation of *Videx* called *Videx EC*® was approved in Europe and the



United States. It is the first single capsule, once-a-day antiretroviral. Similarly, we are developing an extended release formulation of *Zerit*, which we hope will further optimize the conditions for adherence. And there is clear evidence that our protease inhibitor in development is effective when given once-a-day – a first for the protease inhibitor class.

Despite the difficult challenges faced by HIV/AIDS researchers today, we have reason to be confident that we can continue to improve the state of antiretroviral therapy. The articles in this *Nature Insight* capture important advances in our understanding of T-cell memory, molecular mechanisms of viral entry, and HIV immune evasion, among other topics. At Bristol-Myers Squibb we are already putting these and other learnings to good use as we pursue our mission of extending and enhancing the lives of people living with HIV/AIDS.

Outside the laboratory and clinic, Bristol-Myers Squibb has been active for several years in seeking out partnerships for finding innovative and workable approaches to dealing with the HIV/AIDS crisis, particularly in sub-Saharan Africa, the region of the world most affected by the epidemic. In May 1999, the company launched its \$100 million, *SECURE THE FUTURE*™ five-year initiative to fund research, training and community outreach in southern Africa. In May 2000, as part of the ACCESS initiative, the company lowered its prices of AIDS medicines in poor countries by 90 percent of what they cost in the developed world. And in March 2001, the company again did its part to support broad and speedy access to care and therapy by increasing its commitment to *SECURE THE FUTURE* by \$15 million, making the patent to *Zerit* available in South Africa at no cost and lowering the prices of *Zerit* and *Videx* to below cost. Within the framework of the ACCESS initiative, combination therapy of *Zerit* and *Videx* now costs \$1 a day in every country in Africa that wishes to participate in this initiative.

### Peter S. Ringrose, Ph.D.

Chief Scientific Officer, Bristol-Myers Squibb and  
President, Pharmaceutical Research Institute



## Contacts

**Publisher:** Fabien Savenay

**Editor:** Paul Smaglik

**Sales Director:** Ben Crowe

## European Head Office, London

The Macmillan Building  
4 Crinan Street  
London N1 9XW, UK  
Tel +44 (0) 20 7843 4961  
Fax +44 (0) 20 7843 4996  
e-mail: [classified@nature.com](mailto:classified@nature.com)

## Group European Manager:

Leonie Welss (4954)

## European Manager:

Nevin Bayoumi (4978)

## European Sales Consultant:

Jonathan Bull (4973)

## UK/ RoW/ Ireland:

Matt Powell (4953), Ben Corp (4974),

Andy Douglas (4975)

**Holland/ Italy:** Nevin Bayoumi (4978)

**Scandinavia:** Sille Opstrup (4994)

## Production Manager:

Billie Franklin

To send films and materials use

London address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: [classified@nature.com](mailto:classified@nature.com)

## International

## Advertising Coordinator:

Laura Pearson (4977)

## Naturejobs web development:

Tom Hancock

## Naturejobs online production:

Ben Lund

## European Satellite Offices

## France/ Belgium:

Christine Niox-Chateau

Tel + 33 (0) 1 43 20 16 51

Fax + 33 (0) 1 43 20 51 52

e-mail: [c.nioxchateau@nature.com](mailto:c.nioxchateau@nature.com)

## Spain/ Portugal:

Amanda Wren

Tel + 34 91 3913835

Fax + 34 91 319 0091

e-mail: [a.wren@nature.com](mailto:a.wren@nature.com)

## Germany/ Austria/ Switzerland:

Patrick Phelan/ Kate Turner

Tel + 49 89 54 90 57 11/-2

Fax + 49 89 54 90 57 20

e-mails: [p.phelan@nature.com](mailto:p.phelan@nature.com)

[k.turner@nature.com](mailto:k.turner@nature.com)

## US Head Office, New York

345 Park Avenue South,

10th Floor, New York, NY 10010-1707

Tel +1 800 989 7718

Fax +1 800 989 7103

e-mail: [classified@nature.com](mailto:classified@nature.com)

## US Sales Director:

Ben Crowe

## US Sales Manager:

Peyton Mason

## US Advertising Coordinator:

Corissa Salzman

## Japan Head Office, Tokyo

Shin-Mitsuke Building (4F),

3-6 Ichigaya Tamachi Shinjuku-Ku,

Tokyo 162

Tel +81 3 3267 8751

Fax +81 3 3267 8746

e-mail: [k.cowan@nature.com](mailto:k.cowan@nature.com)

## Japan Manager:

Kate Cowan

## Reinventing Naturejobs

Scientific advances plus funding availability equals job opportunities. Of course, several variables conspire to lend that formula its real-life complexity. Naturejobs, in its newly redesigned and expanded format, will strive better to understand and explore these variables and untangle some of the complexities. The section will do this through a number of new and expanded features. This weekly column will look at job prospects in various disciplines, sectors and countries, based on recent news items, conferences, reports or conversations.

**Naturejobs Careers and Recruitment** features will run monthly, rather than six times a year. These features will continue to examine the dynamics of employment in different fields and give a realistic assessment of employment opportunities.

**Our monthly Special Reports** will continue to provide a more focused look at specific projects, programmes or problems. In this issue, we report on the Alliance for Cell Signaling based at the University of Texas Southwestern. This is a notable programme because it represents a direction in which life sciences has been heading since the advent of the Human Genome Project — 'big biology' with multiple centres and millions of dollars.

Another new feature will provide a glimpse of the movement within and between disciplines and sectors. This **Movers** section can by no means capture every recent job appointment. Instead, we aim to take a snapshot of changes that we think represent how scientific employment is changing — and staying the same. Although the individual impressions given in this column, the reporting in the longer features and anecdotes in the **Movers** section may not, in themselves, summarize the entire picture of employment in science, we hope the parts add up to a larger whole.

## Paul Smaglik

Naturejobs editor



## Contents

### SPECIAL REPORT

Alliance seeks members to map cell signalling pathways p4

### MOVERS

Scottish life scientist goes home, physicist splits work between continents, and more Back page

### WWW.NATURE.COM/ NATUREJOBS

Career centre  
Information on the scientific job market

### FOCUS

### SPOTLIGHT

### RECRUITMENT

### SCIENTIFIC ANNOUNCEMENTS

### SCIENTIFIC EVENTS





# Alliance signals a fresh type of scientific research endeavour as the post-genomic face of 'big biology' gets under way. Diane Gershon investigates.

**W**ith seven laboratories, about 50 principal investigators, and over US\$25 million in federal funding for the first 5 of its 10 years, the Alliance for Cell Signaling (AFCS) is serving as an example of the post-genomic face of 'big biology'. This multi-disciplinary effort will focus on the G-protein-regulated (and related) signalling systems in two types of mouse cell — B lymphocyte and specialized heart cells known as cardiac myocytes.

The sheer size of the organization mirrors the complexity of its goals — identifying all relevant signalling molecules in the mouse cells, determining how they are connected, quantifying how information flows through those connections, and building mathematical models that describe how the cellular signalling system works, with the ultimate aim of building a 'virtual cell'. Meeting those goals will require the direct involvement of hundreds of scientists and technicians and indirect input from hundreds more.

Al Gilman, a professor of pharmacology at the University of Texas Southwestern (UT Southwestern) Medical Center in Dallas, began two years ago to assemble an impressive cadre of researchers to work to understand fully how cells interpret signals in a

context-dependent manner. "For years my own research had been going in a more and more reductionist direction but I had always been wondering how one could turn it around," says Gilman, who was awarded the 1994 Nobel prize for his discovery of G proteins and their role in signal transduction in cells.

The alliance became official last September after it was awarded US\$25 million over five years from the US National Institute of General Medical Sciences (NIGMS), part of the National Institutes of Health (NIH). The award was the first in a series of 'glue' grants from the NIGMS for large-scale projects involving multiple institutions (see *Nature* **407**, 7; 2000). Further funds from other NIH institutes, the pharmaceutical industry, non-profit organizations and an anonymous Dallas philanthropist could boost the five-year total to \$50 million, making the AFCS a formidable research force.

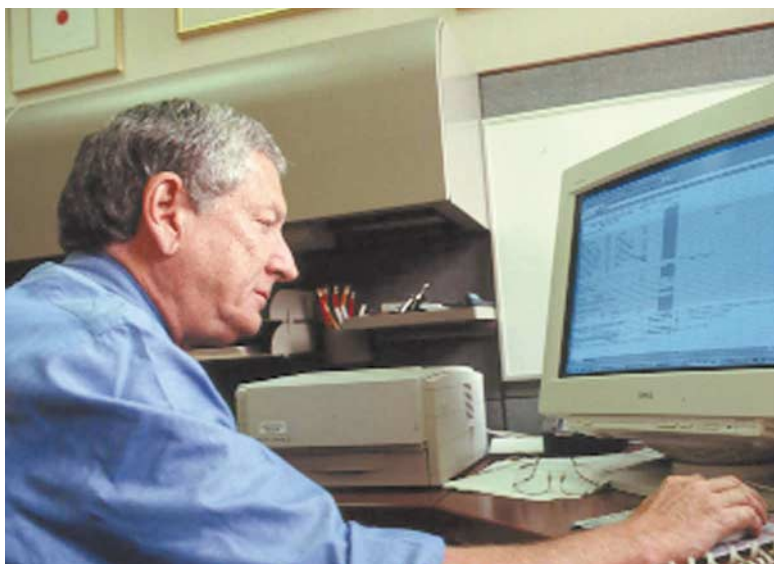
## CELL NETWORKING

Most research related to the project will be carried out in dedicated alliance laboratories staffed by alliance employees. There are currently seven such facilities at five different institutions nationwide — UT Southwestern, the California Institute of Technology, the University of California at San Diego, the San Francisco VA Medical Center (VAMC)/University of California, San Francisco (UCSF) and Stanford University. It is purely coincidental, says Gilman, that all — bar the three labs at UT Southwestern — are on the west coast. For a distributed effort such as this, regular communication is vital and participants stay connected through teleconferencing.

Although the alliance labs are run by lab directors who have agreed to dedicate a percentage of their time to alliance business, scientists recruited specifically by the alliance are not tenure-track and are not allied to a particular department. Roughly 85 of the anticipated 100 people needed to staff the alliance labs are in place; of these, a third will be PhDs.

The jobs run the gamut — from biochemists, cell biologists, microscopists, molecular biologists and protein chemists to programmers and statisticians. A list of job openings can be found on the AFCS website, although some remaining technical spots will not be filled until the project moves into high gear.

**Clear message:** Al Gilman has been instrumental in setting up an alliance for studying how cells interpret signals.



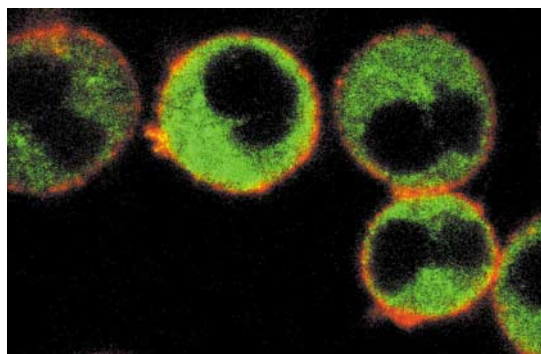
The jobs may not be 'classic' academic positions but they will provide individuals with the opportunity to work with some of the best people in the field and be at the forefront of signalling research. "For people who can't decide whether to go into industry or an academic setting, this is a great hybrid," says Paul Simpson, a professor of cardiology at UCSF and co-director of the alliance lab at the San Francisco VAMC.

#### WANTED: MEMBERS

It is not surprising, given the lack of conventional rewards (most notably publications) with this project, that one of the biggest challenges for the alliance has been in motivating young investigators to become involved. Participating investigators must waive all intellectual property rights to discoveries stemming from alliance-funded work — data that the project generates will be placed in the public domain on the alliance's website.

But some are willing to join anyway. "I'm 60 years old. It's much easier to motivate me because I've had my career. I can take the risk," says Henry Bourne, a professor of medicine at UCSF and chair of the alliance's B-lymphocyte system committee.

In an effort to engage the broader signalling research community, investigators, particularly the younger ones, are encouraged to apply for membership in the alliance by signing up to maintain and prepare a 'molecule page'. The short-form minipages, examples of which can be found on the website, will eventually be superseded by more detailed full-



Under investigation: by researching cell signalling, scientists hope to build a virtual cell to model how systems react.

length molecule pages. These more standardized documents will be vetted by an alliance-appointed editorial board and will contain detailed information about a specific molecule of interest to the alliance. The pages will be posted on the Internet and will form the core of the alliance's signalling data base.

Gilman says he is often asked: 'what's in it for me to do a molecule page?'. He replies: "We'd really like them to be considered to be a worthwhile scholarly document and, hopefully, it will be worth a point or two for the authors when they come up for promotion."

Diane Gershon is assistant editor, new technology, at *Nature Medicine*.

#### Web links

♦ <http://www.cellularsignaling.org>

♦ <http://afcs.swmed.edu>

♦ <http://genomics.lbl.gov>

Paul Simpson says alliance jobs provide 'a great hybrid' between academia and industry.



## Alliance—engineering partnership

Science has long looked down on engineering, says John Doyle, a professor of control and dynamical systems at the California Institute of Technology. Consequently, engineers are not easily persuaded to study biological systems. But, at least in some quarters, there are signs that attitudes may be changing.

Doyle and his collaborator, Adam Arkin, an assistant professor in the department of bioengineering and chemistry at the University of California, Berkeley, have found a very receptive audience in the recently formed Alliance for Cell Signaling. This 10-year, multi-institutional and multidisciplinary programme will study G-protein-mediated and related signalling systems in an effort to glean a more comprehensive understanding of cellular signalling networks.

Although the project will generate a wealth of experimental data over the years, the ultimate goal is to construct a complete virtual representation of cellular signalling — a so-called virtual cell. The hope is that scientists will then be able to make quantitative predictions about how the systems will behave if they are perturbed either by genetic manipulation or by drugs.

As a mark of its commitment to its modelling component, the alliance is involving Doyle and Arkin at the ground level. Doyle, for example, is on the alliance's steering committee and Arkin its B-lymphocyte system committee.

"You can't design experiments and data management systems without thinking about how they're going to be used," says Henry Bourne, a professor of

medicine at the University of California, San Francisco.

#### BUILDING BRIDGES

For now, the alliance's modelling effort is represented by a bridging project. Doyle and Arkin are lead investigators on the project and both are using what are fairly modest levels of alliance support in these early stages as a lever against other sources of funding. In the short run, they are focusing on experimentally more accessible systems than the alliance cell types. This will include developing a sporulation model for the bacterium *Bacillus subtilis*.

The alliance's intention to make systematized measurements using standard protocols, standard cell preparations and standard reagents "is a modeller's dream", says Arkin. But there are theoretical issues to be

overcome. "These are big systems that are hard to measure," he says. "They're highly non-linear and they're noisy."

One issue is that the people who understand large, complex engineering networks are in high demand in the private sector and universities find it hard to compete. They include control theorists, communication theorists and theoretical computer scientists as well as people who can design circuits.

Arkin's senior programmer has 20 years of experience and specializes in the visualization and analysis of multivariate data and multiscale modelling. Arkin says it takes a fairly mature mind and that level of skill and expertise to bring a project like this, which will deal with scales of data ranging from whole-cell morphology to gene expression, to fruition.

D. G.

## MOVERS

### BIOTECHNOLOGY

In moving from Pharmacia Corporation to Ardana Bioscience, **Maureen Lindsay** is returning to both her geographical and intellectual roots. The switch from the global pharmaceutical company to a five-person start-up allows Lindsay to concentrate more on her own research in reproductive health, the company's focus. It also allows her to return to her native Scotland. In Edinburgh, Lindsay serves as Ardana's chief commercial officer, whereas her last position with Pharmacia was as market company president in New Zealand. Ardana was created in July 2000 to commercialize research developed over the past 28 years by the Medical Research Council Human Reproductive Sciences Unit in Edinburgh. The company's next step is to recruit a clinical director.

After nearly 21 years in the United States Navy, **Steve Hoffman** retired from active duty to take a position as senior vice-president at Celera Genomics in Rockville, Maryland. His primary responsibilities will be to use Celera's genomics and proteomics data and bioinformatics capabilities to develop vaccines, therapeutic antibodies, and biologicals against cancer, infectious diseases, and inflammatory and autoimmune diseases. Hoffman is also coordinating the company's interaction with the *Anopheles gambiae* Genome Project International Consortium. Celera's role will be to sequence and assemble the genome of *A. gambiae*, the mosquito considered to be the most important carrier of the malaria parasite *Plasmodium falciparum*.

### PHYSICS

This spring **Elisabeth Giacobino** continues her ascent at the Centre National de la Recherche Scientifique (CNRS) in Paris, with an appointment to director of the CNRS physical sciences and mathematical department. Giacobino joined the CNRS in 1969, working as a researcher in the Hertzian spectroscopy laboratory, which has since become the Kastler-Brossel Laboratory. In 1982 she was promoted to research director and was appointed head of the Kastler-Brossel Laboratory in 1999. Giacobino is studying quantum fluctuations and the utilization of the quantum properties of light for information processing. This work could lead to industrial applications of opto-electronic systems

for sectors such as telecommunications and information technologies. She succeeds Jean-Paul Pouget, who had held this position since October 1997.

**Paul Chu**, who led a group in the discovery of yttrium barium copper oxide, the first material that became superconducting when cooled by liquid nitrogen, will become president of Hong Kong University of Science and Technology starting in July. Chu will maintain his research group at the Texas Center for Superconductivity at the University of Houston, where he has worked since 1987. He looks forward to juggling both responsibilities. "There is a way when there is a will," he says. "To be a president will be my main exciting and challenging job in Hong Kong, and to do research will continue to be my lifetime hobby in Houston."

**Persis Drell**, professor of physics at Cornell University, New York, will lead the Cornell group at CLEO — one of the world's most advanced and sensitive particle detectors. CLEO is used to study elementary particles produced by the Cornell Electron Storage Ring (CESR), an electron-positron collider. Among other things, it is being used to test the standard model, which describes the properties of elementary particles. Cornell's CLEO group has 42 researchers and staff. Drell's priority is to complete and commission the US\$15 million upgrade to the CLEO detector, known as CLEO III, that was started in 1994 and will be finished this year.

### COMPUTER SCIENCE

The long-time director of the University of California's San Diego Supercomputer Center (SDSC) and the National Partnership for Advanced Computational Infrastructure (NPACI), **Sid Karin**, has stepped down to pursue other goals. Computer scientist Francine Berman was elevated to the top post of both SDSC and NPACI, a consortium of research institutions funded by the National Science Foundation. After 16 years at SDSC, Karin said he plans to consider a number of career options, including research, teaching and involvement with start-up companies on whose boards of directors he serves.



Maureen Lindsay



Elisabeth Giacobino



Paul Chu

### CONTACTING US AT MOVERS

Please send any information on new appointments or job moves to: [naturejobseditor@naturedc.com](mailto:naturejobseditor@naturedc.com).